Dear Dr. de Valpine,

Thank you for your insightful and thorough comments. I have addressed each of them below.

Sincerely,

David J. Harris

Dear Mr. Harris,

Thank you very much for submitting your revised manuscript “Estimating species interactions from observational data with Markov networks” (Statistical Report) for review by Ecology. Many aspects of the presentation are improved. Unfortunately my assessment is that the new community dynamics simulations turned out to be a step in the wrong direction, making it harder rather than easier to assess the new method and compare it to older ones. I can see that this new material resulted from taking Reviewer #1’s suggestion to heart, but the bottom line is that it didn’t turn out well.

First, I don’t think you needed to remove the extensive set of potentially very informative simulations that you had done previously but, if anything, add to them. (I know the 20-page Report limit is tight, but you have no limits in the supplement.) Evaluation of new methods when assumptions are not met builds naturally on first having evaluated the method when its assumptions are met, not skipping over that step. However, you have removed the lion’s share of such comparisons, leaving only the simple 3 species case before diving into the community dynamics simulations. I was left unable to know how the method performs on its own terms before being faced with evaluation of how it performs on very complicated terms.

Your proposed community dynamics model would raise a host of concerns. You state it is a large population approximation but then include demographic stochasticity, which is typically of diminishing importance for large populations and outweighed by environmental stochasticity, which you omit. The choice of parameters to draw from is arbitrary and ultimately discernible only by reading code, and the mutualism term appears to be arbitrary. You may disagree, and one could debate all aspects, but that is really the point: one would have wanted a simpler and more standard model so comparison of the statistical methods is not tangled in these issues.

I can see you made a serious effort, but the outcome is not something I think readers will gain value from. The ultimate problem is that we are left with no comprehension of how the data generated by these processes suit the assumptions of the various methods. Does the model actually generate data that are well approximated by the Markov network assumptions or not? Trying to understand that would be a theoretical ecology exercise that, for your proposed model, does make sense to include in a paper about the statistical methods.

I suggest instead staying closer to the statistical models and including some additional cases to probe the method's robustness to violations of its assumptions and/or cases where one of the other parametric models in the competition generates the data. Which kinds of potential violations of assumptions to consider is for you to decide. For example, perhaps the cases of interest for the Markov network model would be if there really are 3-way interactions, or asymmetric interactions, or heterogeneity among sites in the model parameters, that contribute to the data generating process but are ignored in the analysis model. Indeed, describing what would constitute violations would help the reader. Another good way to probe would be to generate data from one of the other parametric models, such as the GLM.

- I appreciate this perspective. I felt very conflicted about including these simulations in the first place, and had many of the same concerns that you raise. I ultimately decided to include them in the previous version because a substantial number of colleagues that read the preprint had similar concerns to Reviewer 1: they wanted to see how the different methods would perform on more realistic data sets (especially on data sets that were not simulated from a model that resembled a Markov network). I do agree, however, that it is important to show that the model performs well on its own terms before diving into a more complicated model, and that omitting the simpler simulations made the paper worse. I had previously considered the option to use a set of GLMs as the generative model (as you suggest), but decided against it once I realized that this approach would either involve directed cycles and the associated identifiability problems (if $\beta_{ij} \neq \beta_{ji}$) or would match the Markov network exactly (if $\beta_{ij} = \beta_{ji}$).

  For this reason, I now include three different simulation regimes with different properties for the 20-species landscapes: 1. Simulating directly from a Markov network, 2. Simulating from a Markov network with spatial heterogeneity in the α parameters (sometimes called a conditional random field), and 3. Simulating from a process that included abundances and per-capita effects rather than species-level interactions between binary species. In order to avoid the problems that you highlighted with the previous submission, I simulated the abundances for the third set of landscapes using Gibbs sampling rather than a population dynamic model. To keep the abundances finite without adding arbitrary constraints to the new simulations, I removed mutualistic interactions from the third set of landscapes.

The overall result is that the reader can now see how the different models perform when A) the species' "true" joint distribution is fully described by univariate and bivariate potentials (e.g. a Markov network), when B) the joint distribution also depends on external environmental factors, and when C) the strength of the interaction between two species depends on (unobserved) variation in abundance.

Note that often violations of assumptions do not create bias in parameter estimates (although sometimes they do) but rather incorrect assessments of uncertainty (inaccurate confidence intervals and Type I error rates in hypothesis tests). You have referred in multiple places to having applied a bootstrap as well as calculating Wald intervals, but it appears to me you have reported almost no results about validity of confidence intervals or hypothesis tests from those (there are a few results for the 3-species case). It would be important to do so.

- I now report approximate Type I error rates and the coverage of the model's 95% confidence intervals. The 95% confidence intervals have about 98% coverage when the data was generated from an unmodified Markov network, and about 86% coverage in the presence of environmental heterogeneity (lines 201-206 and Appendix 4). These coverage values aren't perfect, but they are reasonable, and they provide far better inferences than ecologists' current methods (lines 206-213; Figure 4C).

It appears to me that you did not understand my point that you are presenting results that are integrated over parameter distributions, which appears still to be the case. What I meant is the following. Often one would want to see how a statistical method performs as a function of specific parameter values. For example, one might vary the value of an interaction coefficient (beta) from negative (competition) to zero to positive (mutualism) and run a set of simulations for each value of beta. This would allow evaluation of statistical power and accuracy of confidence intervals and Type I error rate as well as RMSE, all as a function of true parameter values. It would also allow assessment of whether the method works better for some parameter scenarios than others, which is not uncommon. Instead you are drawing parameters from distributions and then presenting performance assessments in aggregate from those results, such as the average R-squared over all cases of a given network size. Generally, and clearly in your case, such metrics can be written as expectations (integrations) over the distribution of parameters you sampled from. You are right there is nothing Bayesian about it, but I pointed out a Bayesian may feel justified in doing it because many Bayesian results take the same format. It is a question of presenting aggregated versus disaggregated results. Unfortunately, presenting only aggregated results leaves the reader unable to dig deeper into the disaggregated results. The analogy to sampling from an archipelago is lovely but irrelevant. It appears to me that since you have all the simulations at hand, and since you can place more detailed results in the supplement, you should do so. I leave it to you to decide how to do that.

- Showing the results' dependence on the value of beta is a great idea, and I have added several analyses along these lines in Figure 4C and Appendix 4. In particular, I now show how CI coverage, RMSE, and the probability that the model will predict the wrong sign for an interaction vary as a function of beta (in addition to global estimates of R-squared and Type I error rates). If you think that readers would benefit from even more disaggregation, I could include a data file containing the seven models' estimates for all the species pairs across all 450 simulated landscapes as a fifth Appendix.

Beyond these steps, I've considered several other options for further exploration of the parameter space, but none of them would be very useful given the large number of parameters in these models (20 alphas and 190 betas). Systematically varying one of the beta coefficients while holding the other 209 parameters constant at arbitrary values would not convincingly yield generalizable insights, as different lines through the 210-dimensional space could produce different results.

Some other minor notes

I don't think you have stated that the betas are symmetric. It appears that beta_ij = beta_ji and hence each could be interpreted as half of the interaction strength. E.g. on line 80, it would be e^4 if I take the preceding model equation as correct and include both beta_ij and beta_ji. Please fix it up if that is not right.

- I've made this issue much clearer in the revision: see lines 68-78. It should now be unambiguous that the sum in the likelihood should only include each pair once, and that each pair must share a single value for β_{ij}.

Both times that I have seen the title anew, the phrase "estimating species interactions from observational data" made me think it would be about time-series method, since that is a common kind of "observational data" from which people try to estimate interactions. I recommend adding a descriptive term like "species composition data" or "observational occurrence data" or "presence/absence data." You decide. This could enter the first Abstract sentence as well.

- This is another good point. I've changed the title to focus on co-occurrence data. Now that I've included inferential statistics in the paper and evaluated their frequentist properties, I've also changed the word "estimating" back to "inferring".

Line 44 is not stated well. You could replace "begin with the assumption" by "test the hypothesis" or insert "that the hypothesis of interest is that [all pairwise interactions…]" or some such change.

- I have changed the wording to emphasize hypothesis testing rather than assumptions (lines 40-42).

Line 69: The phrase "how groups of species can co-occur" seems imprecise or incorrect. Maybe "to determine species occurence models"?

- Lines 65-66 now describe "how… conditional relationships can determine species composition" rather than referring to "groups".

Lines 86-87 are not a good lead-in to the point of the rest of this paragraph. It sounds like you are about to talk about species interactions, again, but really the point of the paragraph is that normalizing the probabilities involves a difficult summation.

- The discussion of computational difficulties has been removed to save space, and the lines you noted are now at the end of a paragraph rather than the beginning of one (lines 82-85).

Lines 72 & 101. I was willing to overlook the fact that you didn't define the vector y explicitly in terms of its elements on line 72 because perhaps that is really obvious. However on line 101 you then state maximum likelihood as if it would be done from a single y vector (and a single likelihood term) rather than a matrix of observations. So now I think you need more explicit notation to clarify the arrangements of your variables.

- The revised manuscript now clarifies that each $\vec{y}$ vector represents the different species that could occur at a single site and that the full co-occurrence matrix is made of "a set of independent $\vec{y}$ vectors indicating which species are present at each site on the landscape" (lines 89-90). This should clarify the relationship between the $\vec{y}$ vectors and the full matrix as well as the relationship between the single-site likelihood of Equation 1 and the landscape-level likelihood (i.e. independence among sites implies that the full likelihood equals the sum of the site-level likelihoods).

Line 130-132. This statement is unclear and unconvincing. Are you saying that the remoteness of the upper range of the prior justifies that the prior has no influence? That would not be correct since it is an issue of prior weightings in the range of heavy posterior weight. The simplest way to show a prior has little influence is to try a couple of different ones.

- I no longer suggest that the prior has no influence. I also make no claims (e.g. regarding unbiasedness) that would be invalidated by a non-flat prior, so the strength of the prior should not affect the validity of any of the results presented. Your initial concern about the prior in my first submission was that the logistic distribution with scale 1 was too similar to the exponential distribution with rate 1 used to simulate the "true" distribution. Now that the prior is more than twice as wide as the "true" distribution, this is no longer an issue.

  With regard to your suggestion that I try different priors, I have now used three different ones and gotten similar results: a flat prior (for the three-species landscapes where the MLE was always finite), and two different logistic priors (for the larger communities).

Lines 136-137. Bootstrapping is a method to estimate uncertainty of an estimation procedure. It is unclear what you mean by using it to validate if a procedure gives “stable” estimates.

- The bootstrapping results have now been removed and replaced with the approximate confidence intervals.

Line 154: This statement is not clear.

- See lines 144-148 for a clarified version

Figure 2B: This is odd notation because, if I follow, you have chosen a null symbol (0 with a slash) to indicate the event y_i = 0 but have chosen the label y_i to indicate the event y_i = 1. I think you want P[00], P[10], etc. or P[y1=0, y2=0], P[y1=1, y2=0], etc, or something like that. The caption states “relative probability” where I think you want “probability.”

- This is an excellent point. The figure has been fixed (along with the caption).

Figures in general: would benefit from titles on each subfigure.

- Subfigures now have titles

It’s too bad we are in this unusual situation that your effort to respond to a reviewer made the manuscript worse. I will let you make another revision to get back on course. What I need you to do are: (1) remove the community dynamics model, put back in extensive results from the model of interest, and take a new approach to probing the model more thoroughly by including some scenarios with clearly defined violations of assumptions and/or data generated from one of the alternative statistical models. You may decide what to put in the online material since I recognize you are trying to stick to the 20-page report limit.

- As described above, I have removed the community dynamics model and now compare all the models' performance on simulated landscapes where all the assumptions are met, and in two cases where different assumptions are violated in clear ways.

2 Include meaningful results about accuracy of uncertainties. These should be at your fingertips since you can calculate confidence intervals for every simulation.

- As described above, I now estimate the coverage of the confidence intervals, and (Appendix 4) show how coverage varies as a function of the "true" parameter value.

3 Provide some disaggregated results.

- As described above, I now show how several performance metrics vary as a function of β and have offered to include the raw results for all 7 methods across the 450 simulations.

4 Show how the regression steps work in about a page or two of supplement text, not requiring code reading.

- I wasn't entirely sure whether you were referring to the logistic regression models or to the process of fitting the Markov network, but I believe that I've clarified the discussion of both issues sufficiently that supplemental material is no longer required for either:

  With regard to the logistic regressions, I've now clarified the symmetry issue: the reader should now be able to tell that I am merely averaging the estimated coefficient from the regression of y_i on y_j with the coefficient from the regression of y_j on y_i (lines 142-148).

With regard to fitting the Markov network, the function being optimized is already in Equation 1 and its gradients are clearly derived in Murphy (2012; see pages 676-677 in the linked PDF from author's website). The manuscript text now clarifies that the role of the package is simply to define these equations in R code and to pass them to R's general purpose optimizer (lines 90-98).