Title: Estimating species interactions from observational data with Markov networks

Author: David J. Harris: Population Biology; 1 Shields Avenue, Davis CA, 95616

Abstract

Inferring species interactions from observational data is one of the most controversial tasks in community ecology. One difficulty is that a single pairwise interaction can ripple through an ecological network and produce surprising indirect consequences. For example, two competing species would ordinarily correlate negatively in space, but this effect can be reversed in the presence of a third species that is capable of outcompeting both of them when it is present. Here, I apply models from statistical physics, called Markov networks or Markov random fields, that can predict the direct and indirect consequences of any possible species interaction matrix. Interactions in these models can be estimated from observational data via maximum likelihood. Using simulated landscapes with known pairwise interaction strengths, I evaluated Markov networks and several existing approaches. The Markov networks consistently outperformed other methods, correctly isolating direct interactions between species pairs even when indirect interactions or abiotic environmental effects largely overpowered them. A linear approximation, based on partial covariances, also performed well as long as the number of sampled locations exceeded the number of species in the data. Indirect effects reliably caused a common null modeling approach to produce incorrect inferences, however.

Key words: Ecological interactions; Occurrence data; Species associations; Markov network; Markov random field; Ising model; Biogeography; Presence–absence matrix; Null model

Introduction

Ecologists' intense interest in drawing inferences about species interactions —especially competition—from presence-absence data has a long history (MacArthur 1958, Diamond 1975, Connor et al. 2013). If nontrophic species interactions are important drivers of community assembly patterns, then we should expect to see their influence in our data sets. Despite decades of work and several major controversies, however (Lewin 1983, Strong et al. 1984, Gotelli and Entsminger 2003), existing methods for detecting competition's effects on community structure are unreliable (Gotelli and Ulrich 2009). More generally, it can be difficult to reason about the complex web of direct and indirect interactions in real assemblages, especially when these interactions occur against a background of other ecological processes such as dispersal and environmental filtering (Connor et al. 2013). For this reason, it isn't always clear what kinds of patterns would even constitute evidence of competition, as opposed to some other biological process or random sampling error (Lewin 1983, Roughgarden 1983).

Most existing methods in this field compare the frequency with which two putative competitors are observed to co-occur against the frequency that would be expected if *all* species on the landscape were independent (Strong et al. 1984, Gotelli and Ulrich 2009). Examining a species pair against such a "null" background, however, rules out the possibility that the overall association between two species could be driven by an outside force. For example, even though the two shrub species in Figure 1 compete with one another for resources at a mechanistic level, they end up clustering together on the landscape because they both grow best in areas that are not overshadowed by trees. If this sort of effect is common, then significant deviations from independence will not—by themselves—provide convincing evidence of species' direct effects on one another.

While the competition between the two shrubs in the previous example does not leave the commonly-expected pattern in community structure (negative association at the landscape level), it nevertheless does leave a signal in the data (Figure 1C). Specifically, *among shaded sites*, there will be a deficit of co-occurrences, and *among unshaded sites*, there will also be such a deficit.

In this paper, I introduce Markov networks (undirected graphical models also known as Markov random fields; Murphy 2012) as a framework for understanding the landscape-level consequences of pairwise species interactions, and for detecting them with observational data. Markov networks, which generalize partial correlations to non-Gaussian data, have been used in many scientific fields to model associations between various kinds of "particles". For example, a well-studied network called the Ising model has played an important role in our understanding of physics (where nearby particles tend to align magnetically with one another; Cipra 1987). In spatial contexts, these models have been used to describe interactions between adjacent grid cells (Harris 1974, Gelfand et al. 2005). In neurobiology, they have helped researchers determine which neurons are connected to one another by modeling the structure in their firing patterns (Schneidman et al. 2006). Following recent work by Azaele et al. (2010) and Fort (2013), I suggest that ecologists could similarly treat species as the interacting particles in this modeling framework. Doing so would allow ecologists to simulate and study the landscape-level consequences of arbitrary species interaction matrices, even when our observations are not Gaussian. While ecologists explored some related approaches in the 1980's (Whittam and Siegel-Causey 1981), computational limitations had previously imposed severe approximations that produced unintelligible results (e.g. "probabilities" greater than one; Gilpin and Diamond 1982). Now that it is computationally feasible to fit these models exactly, the approach has become worth a second look.

The rest of the paper proceeds as follows. First, I discuss how Markov networks work and how they can be used to simulate landscape-level data and to predict the direct and indirect consequences of possible interaction matrices. Then, using simulated data sets where the "true" ecological structure is known, I compare this approach with several existing methods for detecting species interactions. Finally, I discuss opportunities for extending the approach presented here to larger problems in community ecology.

Methods

Conditional relationships and Markov networks. Ecologists are often interested in inferring direct interactions between species, controlling for the indirect influence of other species. In statistical terms, this implies that ecologists want to estimate *conditional* ("all-else-equal") relationships, rather than *marginal* ("overall") relationships. The most familiar conditional relationship is the partial correlation, which indicates the portion of the sample correlation between two species that remains after controlling for other variables in the data set (Albrecht and Gotelli 2001). The example with the shrubs and trees in Figure 1 shows how the two correlation measures can have opposite signs, and suggests that the partial correlation is more relevant for drawing inferences about species interactions (e.g. competition). Markov networks extend this approach to non-Gaussian data, much as generalized linear models do for linear regression (Lee and Hastie 2012).

Markov networks give a probability value for every possible combination of presences and absences in communities. For example, given a network with binary outcomes (i.e. 0 for absence and 1 for presence), the relative probability of observing a given presence-absence vector, $\vec{y}$, is given by

$$p(\vec{y}; \alpha, \beta) \propto exp(\sum_i \alpha_i y_i + \sum_{i \neq j} \beta_{ij} y_i y_j).$$

Here, α_i is the amount that the presence of species i contributes to the log-probability of $\vec{y}$; it directly controls the prevalence of species i. Similarly, β_{ij} is the amount that the co-occurrence of species i and species j contributes to the log-probability, and controls how often the two species will be found together (Figure 2A, Figure 2B). β thus acts as an analog of the partial covariance, but for non-Gaussian networks. Because the relative probability of a presence-absence vector increases when positively-associated species co-occur and decreases when negatively-associated species co-occur, the model tends to produce assemblages that have many pairs of positively-associated species and relatively few pairs of negatively-associated species (exactly as an ecologist might expect).

A major benefit of Markov networks is the fact that the conditional relationships between species can be read directly off the matrix of β coefficients (Murphy 2012). For example, if the coefficient linking two mutualist species is +2, then—all else equal—the odds of observing either species increase by a factor of e^2 when its partner is present (Murphy 2012). Of course, if all else is *not* equal (e.g. Figure 1, where the presence of one competitor is associated with release from another competitor), then species' marginal association rates can differ from this expectation. For this reason, it is important to consider how coefficients' effects propagate through the network, as discussed below.

Estimating the marginal relationships predicted by a Markov network is more difficult than estimating conditional relationships, because doing so requires absolute probability estimates. Turning the relative probability given by Equation 1 into an absolute probability entails scaling by a *partition function*, $Z(\alpha, \beta)$, which ensures that the probabilities of all possible assemblages that could be produced by the model sum to one (bottom of Figure 2B). Calculating $Z(\alpha, \beta)$ exactly, as is done in this paper, quickly becomes infeasible as the number of species increases: with 2^N possible assemblages of N species the number of bookkeeping operations required for exact inference quickly spirals exponentially into the billions. Numerous techniques are available for working with Markov networks that keep the computations tractable, either through analytic approximations (Lee and Hastie 2012) or Monte Carlo sampling (Salakhutdinov 2008), but they are beyond the scope of this paper.

Simulations. In order to compare different methods for drawing inferences from observational data, I simulated two sets of landscapes using known parameters.

The first set of simulated landscapes included the three competing species shown in Figure 1. For each of 1000 replicates, I generated a landscape with 100 sites by sampling exactly from a probability distribution defined by the interaction coefficients in that figure (Appendix A). Each of the methods described below (a Markov network, two correlation-based methods and a null modeling approach) was then evaluated on its ability to correctly infer that the two shrub species competed with one another, despite their frequent co-occurrence.

I also simulated a second set of landscapes with five, ten, or twenty potentially-interacting species on landscapes composed of 20, 100, 500, or 2500 observed communities (24 replicate simulations for each combination; Appendix B). These simulated data sets span the range from small, single-observer data sets to large collaborative efforts such as the North American Breeding Bird Survey. As described in Appendix B, I randomly drew the "true" coefficient values for each replicate so that most species pairs interacted negligibly, a few pairs interacted very strongly, and competition was three times more common than facilitation. I then used Gibbs sampling to randomly generate landscapes with varying numbers of species and sites via Markov chain Monte Carlo (Appendix B). For half of the simulated landscapes, I treated each species' α coefficient as a constant, as described above. For the other half, I treated the α coefficients as linear functions of two abiotic environmental factors that varied from location to location across the landscape (Appendix B). The latter set of simulated landscapes provide an important test of the methods' ability to distinguish co-occurrence patterns that were generated from pairwise interactions among the observed species from those that were generated by external forces like abiotic environmental filtering. This task was made especially difficult because—as with most analyses of presence-absence data for co-occurrence patterns—the inference procedure did not have access to any information about the environmental or spatial variables that helped shape the landscape (cf Connor et al. 2013, Blois et al. 2014).

Inferring α and β coefficients from presence-absence data. The previous sections involved known values of α and β. In practice, ecologists will often need to estimate these parameters from data instead. When the number of species is reasonably small, one can compute exact maximum likelihood estimates for all of the α and β coefficients by optimizing $p(\vec{y}; \alpha, \beta)$. Fully-observed Markov networks like the ones considered here have unimodal likelihood surfaces (Murphy 2012), ensuring that this procedure will always converge on the global maximum. This maximum is the unique combination of α and β coefficients that would be expected to produce exactly the observed co-occurrence frequencies. For the analyses in this paper, I used the *rosalia* package (Harris 2015a) for the R programming language (R Core Team 2015) to define the objective function and gradient as R code. The rosalia package then uses the `BFGS` method in R's `optim` function to find the best values for α and β.

For analyses with 5 or more species, I made a small modification to the maximum likelihood procedure described above. Given the large number of parameters associated with some of the networks to be estimated, I regularized the likelihood using a logistic prior distribution (Gelman et al. 2008) with a scale of 1 on the α and β terms.

Other inference techniques for comparison. After fitting Markov networks to the simulated landscapes described above, I used several other techniques for inferring the sign and strength of marginal associations between pairs of species (Appendix B).

The first two interaction measures were the simple and partial covariances between each pair of species' data vectors on the landscape (Albrecht and Gotelli 2001). Because partial covariances are undefined for landscapes with perfectly-correlated species pairs, I used a regularized estimate based on ridge regression [Wieringen and Peeters (2014); i.e. linear regression with a Gaussian prior]. For these analyses, I set the ridge parameter to 0.2 divided by the number of sites on the landscape.

The third method, described in Gotelli and Ulrich (2009), involved simulating possible landscapes from a null model that retains the row and column sums of the original matrix (Strong et al. 1984). Using the default options in the Pairs software described in Gotelli and Ulrich (2009), I simulated the null distribution of scaled C-scores (a test statistic describing the number of *non*-co-occurrences between two species). The software then calculated a Z-statistic for each species pair using this null distribution. After multiplying this statistic by -1 so that positive values corresponded to facilitation and negative values corresponded to competition, I used it as another estimate of species interactions.

Method evaluation. For the first simulated landscape (three species), I kept the evaluation simple and qualitative: any method that reliably determined that the two shrub species were negatively associated passed; other methods failed.

For the larger landscapes, I rescaled the four methods' estimates using linear regression through the origin so that they all had a consistent interpretation. In each case, I regressed the "true" β coefficient for each species pair against the model's estimate, re-weighting the pairs so that each landscape contributed equally to the rescaled estimate[1]. For each estimate of a species pair's interactions, I used this regression to calculate the squared error associated with method that produced it. Finally, I averaged these squared errors for each combination of species richness, landscape size, statistical method, and presence/absence of environmental filtering across all 12 replicates; the mean squared errors associated with these subsets of the data determined the proportion of variance explained by each method under different conditions.

Results

Three species. As shown in Figure 1, the marginal relationship between the two shrub species was positive—despite their competition for space at a mechanistic level— due to indirect effects of the dominant tree species. As a result, the covariance method falsely reported positive associations 94% of the time and the randomization-based null model falsely reported such associations 100% of the time. The two methods for evaluating conditional relationships (Markov networks and partial covariances), however, successfully controlled for the indirect pathway via the tree species and each correctly identified the direct negative interaction between the shrubs 94% of the time.

Larger landscapes. The accuracy of the four evaluated methods varied substantially, depending on the parameters that produced the simulated communities (Figure 3). In

[1] The null model generated one Z-score outlier greater than 1000, which dominated the regression and squared error analyses. To reduce its influence on these results, I changed its value to 32.5, which was the value of the next largest Z-score in the null model's results.

general, however, there was a consistent ordering: the Markov network explained 54% of the variance overall, followed by partial covariances (32%), sample covariances (21%), and Z scores from the null model (17%).

The models' accuracies tended to decline when environmental filters were added, particularly when the number of species was small and the effects could not be diluted among many pairwise interactions.

Discussion

The results presented above are very promising, as they show that Markov networks can recover species' pairwise interactions from observational data, even when direct interactions are largely overwhelmed by indirect effects (e.g. Figure 1) or environmental effects (lower panels of Figure 3). For cases where it is infeasible to fit a Markov network, these results also indicate that partial covariances—which can be computed straightforwardly by linear regression—can often provide an accurate approximation.

Apart from the environmental filters, the simulated landscapes presented here represent the best-case scenario for these methods. Future research should thus examine these models' performance characteristics when the "true" interaction matrices include guild structure or trophic levels, which could make the β coefficients much more difficult to infer (particularly for linear approximations like the partial covariance approach; Loh and Wainwright (2013)). On the other hand, ecologists may often have prior information about the nature of real species' interaction patterns from natural history or ecological experiments, which could substantially reduce the probability and magnitude of error. The rosalia package (Harris 2015a) has built-in mechanisms for incorporating this kind of information, if it can be expressed as a prior probability distribution or a penalty on the likelihood.

Additionally, it is important to note that, while partial correlations and Markov networks both prevent us from mistaking marginal associations for conditional ones, they cannot tell us the underlying biological mechanism. Real species co-occurrence patterns will depend on a number of factors—especially in taxa that emigrate in response to other species—and the β coefficients in Markov networks have to reduce this to a single number. Thus, experiments and natural history knowledge will generally be required to pin down the exact nature of the interaction (e.g. who outcompetes whom).

Despite these limitations, the results with environmental filtering seem to indicate that the method can be very robust. Additionally, the fact that Markov networks provide a likelihood function to optimize makes them highly extensible, even when it is inconvenient to compute the likelihood exactly. For example, the mistnet software package for joint species distribution modeling (Harris 2015b) can fit *approximate* Markov networks to large species assemblages (>300 species) while simultaneously modeling each species' response to the abiotic environment with complex, nonlinear functions. This sort of approach, which combines multiple ecological processes, could help ecologists to disentangle different factors behind the co-occurrence patterns we observe in nature. Numerous other extensions are possible: similar networks can be fit with continuous variables, count data, or both (Lee and Hastie 2012). There are even methods (Whittam and Siegel-Causey 1981, Tjelmeland and Besag 1998) that would allow the coefficient linking two species in an interaction matrix to vary as a function of the abiotic environment or of third-party species that could tip the balance between facilitation and exploitation (Bruno et al. 2003). Fully exploring these possibilities will require more research into the various available approximations to the log-likelihood and to its gradient, in order to balance efficiency, accuracy, and the ability to generate confidence limits for statistical inference.

By providing precise quantitative expectations about the results of species interactions, Markov networks have the potential for addressing long-standing ecological questions. For example, Markov networks can provide a precise answer to the question of how competition affects species' overall prevalence, which was a major flash point for the null model debates in the 1980's (Strong et al. 1984). From Equation 1, one can derive the expected prevalence of a species in the absence of biotic influences ($\frac{1}{1+e^{-\alpha}}$). Any significant difference between this value and the observed prevalence can be attributed to the β coefficients linking this species to its facilitators and competitors (cf Figure 2D).

This paper only scratches the surface of what Markov networks can do for ecology. This family of models—particularly the Ising model for binary networks—has been extremely well-studied in statistical physics for nearly a century, and the models' properties, capabilities, and limits are well- understood in a huge range of applications, from spatial modeling (Gelfand et al. 2005) to neuroscience (Schneidman et al. 2006) to models of human behavior (Lee et al. 2013). Modeling species interactions using the same framework would thus allow ecologists to tap into an enormous set of existing discoveries and techniques for dealing with indirect effects, stability, and alternative stable states.

These results also have important implications for the continued use of fixed-fixed null models in ecology. The small simulated landscapes described by Figure 1 show that test statistics based on marginal co-occurrence (such as C-scores) will not always have a straightforward relationship with the underlying ecological processes. Moreover, the larger communities analyzed in Figure 3 often fell so far outside the null distribution[2] that it probably makes more sense to reject whole model rather than to assign blame for the

[2] Nearly 20% of the species pairs fell outside the 99.99994% confidence intervals implied by their Z scores ($|Z| > 5$), and about 10% had uncorrected p-values below R's default numerical precision of 2×10^{-16}.

discrepancy to any one species pair. On average, the pairwise Z scores from the null model provided less information about direct species interactions than correlation coefficients did. Researchers using null modeling approaches may be able to predict twice as much of the variance in species' "true" interaction strengths using partial covariances from linear regression, or triple them using a Markov network.

Null and neutral models can be very useful for clarifying our thinking about the numerical consequences of species' richness and abundance patterns (Harris et al. 2011, Xiao et al. 2015), but deviations from a null model must be interpreted with care (Roughgarden 1983). In complex networks of ecological interactions—and even in small networks with three species—it may simply not be possible to implicate individual species pairs or specific ecological processes like competition by rejecting a general- purpose null. Direct estimates of species' conditional associations may be the only way to make these inferences reliably.

Acknowledgements: This research was funded by a Graduate Research Fellowship from the US National Science Foundation and benefited greatly from discussions with A. Sih, M. L. Baskett, R. McElreath, R. J. Hijmans, A. C. Perry, and C. S. Tysor. Additionally, A. K. Barner, E. Baldridge, E. P. White, D. Li, D. L. Miller, N. Golding, and N. J. Gotelli provided useful feedback on an earlier draft of this work.

References:

Albrecht, M., and N. J. Gotelli. 2001. Spatial and temporal niche partitioning in grassland ants. Oecologia 126:134–141.

Azaele, S., R. Muneepeerakul, A. Rinaldo, and I. Rodriguez-Iturbe. 2010. Inferring plant ecosystem organization from species occurrences. Journal of theoretical biology 262:323–329.

Blois, J. L., N. J. Gotelli, A. K. Behrensmeyer, J. T. Faith, S. K. Lyons, J. W. Williams, K.

L. Amatangelo, A. Bercovici, A. Du, J. T. Eronen, and others. 2014. A framework for evaluating the influence of climate, dispersal limitation, and biotic interactions using fossil pollen associations across the late Quaternary. Ecography 37:1095–1108.

Bruno, J. F., J. J. Stachowicz, and M. D. Bertness. 2003. Inclusion of facilitation into ecological theory. Trends in Ecology & Evolution 18:119–125.

Cipra, B. A. 1987. An introduction to the Ising model. American Mathematical Monthly 94:937–959.

Connor, E. F., M. D. Collins, and D. Simberloff. 2013. The checkered history of checkerboard distributions. Ecology 94:2403–2414.

Diamond, J. M. 1975. The island dilemma: Lessons of modern biogeographic studies for the design of natural reserves. Biological conservation 7:129–146.

Fort, H. 2013. Statistical Mechanics Ideas and Techniques Applied to Selected Problems in Ecology. Entropy 15:5237–5276.

Gelfand, A. E., A. M. Schmidt, S. Wu, J. A. Silander, A. Latimer, and A. G. Rebelo. 2005. Modelling species diversity through species level hierarchical modelling. Journal of the Royal Statistical Society: Series C (Applied Statistics) 54:1–20.

Gelman, A., A. Jakulin, M. G. Pittau, and Y.-S. Su. 2008. A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models. The Annals of Applied Statistics 2:1360–1383.

Gilpin, M. E., and J. M. Diamond. 1982. Factors contributing to non-randomness in species Co-occurrences on Islands. Oecologia 52:75–84.

Gotelli, N. J., and G. L. Entsminger. 2003. Swap algorithms in null model analysis. Ecology:532–535.

Gotelli, N. J., and W. Ulrich. 2009. The empirical Bayes approach as a tool to identify non-random species associations. Oecologia 162:463–477.

Harris, D. J. 2015a. Rosalia: Exact inference for small binary Markov networks. R package version 0.1.0. Zenodo. http://dx.doi.org/10.5281/zenodo.17808.

Harris, D. J. 2015b. Generating realistic assemblages with a Joint Species Distribution Model. Methods in Ecology and Evolution.

Harris, D. J., K. G. Smith, and P. J. Hanly. 2011. Occupancy is nine-tenths of the law: Occupancy rates determine the homogenizing and differentiating effects of exotic species. The American naturalist 177:535.

Harris, T. E. 1974. Contact Interactions on a Lattice. The Annals of Probability 2:969–988.

Lee, E. D., C. P. Broedersz, and W. Bialek. 2013. Statistical mechanics of the US Supreme Court. arXiv:1306.5004 [cond-mat, physics:physics, q-bio].

Lee, J. D., and T. J. Hastie. 2012, May. Learning Mixed Graphical Models.

Lewin, R. 1983. Santa Rosalia Was a Goat. Science 221:636–639.

Loh, P.-L., and M. J. Wainwright. 2013. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. The Annals of Statistics 41:3022–3049.

MacArthur, R. H. 1958. Population ecology of some warblers of northeastern coniferous forests. Ecology 39:599–619.

Murphy, K. P. 2012. Machine Learning: A Probabilistic Perspective. The MIT Press.

R Core Team. 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Roughgarden, J. 1983. Competition and Theory in Community Ecology. The American

Naturalist 122:583–601.

Salakhutdinov, R. 2008. Learning and evaluating Boltzmann machines. Technical Report UTML TR 2008-002, Department of Computer Science, University of Toronto, Dept. of Computer Science, University of Toronto.

Schneidman, E., M. J. Berry, R. Segev, and W. Bialek. 2006. Weak pairwise correlations imply strongly correlated network states in a neural population. Nature 440:1007–1012.

Strong, D. R., D. Simberloff, L. G. Abele, and A. B. Thistle. 1984. Ecological communities: Conceptual issues and the evidence. Princeton University Press.

Tjelmeland, H., and J. Besag. 1998. Markov Random Fields with Higher-order Interactions. Scandinavian Journal of Statistics 25:415–433.

Whittam, T. S., and D. Siegel-Causey. 1981. Species Interactions and Community Structure in Alaskan Seabird Colonies. Ecology 62:1515–1524.

Wieringen, W. N. van, and C. F. Peeters. 2014. Ridge Estimation of Inverse Covariance Matrices from High-Dimensional Data. arXiv preprint arXiv:1403.0904.

Xiao, X., D. J. McGlinn, and E. P. White. 2015. A strong test of the Maximum Entropy Theory of Ecology. The American Naturalist 185:E70–E80.

Figure captions

Figure 1. **A.** A small network of three competing species. The tree (top) tends not to co-occur with either of the two shrub species, as indicated by the strongly negative coefficient linking them. The two shrub species also compete with one another, as indicated by their negative coefficient (circled), but this effect is substantially weaker. **B.** In spite of the competitive interactions between the two shrub species, their shared tendency to occur in locations without trees makes their occurrence vectors positively correlated (circled). **C.**

Controlling for the tree species' presence with a conditional method such as a partial covariance or a Markov network allows us to correctly identify the negative interaction between these two species (circled).

Figure 2. **A.** A small Markov network with two species. The depicted abiotic environment favors the occurrence of both species ($\alpha > 0$), particularly species 2 ($\alpha_2 > \alpha_1$). The negative β coefficient linking these two species implies that they co-occur less than expected under independence. **B.** Relative probabilities of all four possible presence-absence combinations for Species 1 and Species 2. The exponent includes α_1 whenever Species 1 is present ($y_1 = 1$), but not when it is absent ($y_1 = 0$). Similarly, the exponent includes α_2 only when species 2 is present ($y_2 = 1$), and β only when both are present ($y_1 y_2 = 1$). The normalizing constant Z, ensures that the four relative probabilities sum to 1. In this case, Z is about 18.5. **C.** Using the probabilities, we can find the expected frequencies of all possible co-occurrence patterns between the two species of interest. **D.** If β equaled zero (e.g. if the species no longer competed for the same resources), then the reduction in competition would allow each species to increase its occurrence rate and the deficit of co-occurrences would be eliminated.

Figure 3. Proportion of variance in interaction coefficients explained by each method with 5, 10, or 20 species arrayed across varying numbers of sampled locations when environmental filtering was absent (top row) or present (bottom row). A negative R^2 values implies that the squared error associated with the corresponding subset of the predictions was larger than the error one would get from assuming that all coefficients equalled zero.