2 **Title: Estimating species interactions from observational data with Markov**

3 **networks**

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

5 ***Abstract***

6 Inferring species interactions from observational data is one of the most controversial tasks in

7 community ecology. One difficulty is that a single pairwise interaction can ripple through an

8 ecological network and produce surprising indirect consequences. For example, two

9 competing species would ordinarily correlate negatively in space, but this effect can be

10 reversed in the presence of a third species that is capable of outcompeting both of them

11 when it is present. Here, I apply models from statistical physics, called Markov networks or

12 Markov random fields, that can predict the direct and indirect consequences of any possible

13 species interaction matrix. Interactions in these models can be estimated from observational

14 data via maximum likelihood. Using simulated landscapes with known pairwise interaction

15 strengths, I evaluated Markov networks and several existing approaches. The Markov

16 networks consistently outperformed other methods, correctly isolating direct interactions

17 between species pairs even when indirect interactions or abiotic environmental effects largely

18 overpowered them. A linear approximation, based on partial covariances, also performed well

19 as long as the number of sampled locations exceeded the number of species in the data.

20 Indirect effects reliably caused a common null modeling approach to produce incorrect

21 inferences, however.

22 **Key words:** Ecological interactions; Occurrence data; Species associations; Markov network;

23 Markov random field; Ising model; Biogeography; Presence–absence matrix; Null model

Introduction

If nontrophic species interactions, such as competition, are important drivers of community assembly, then ecologists might expect to see their influence in our data sets (MacArthur 1958, Diamond 1975). Despite decades of work and several major controversies, however (Lewin 1983, Strong et al. 1984, Gotelli and Entsminger 2003, Connor et al. 2013), existing methods for detecting competition's effects on community structure are unreliable (Gotelli and Ulrich 2009), and thus important ecological processes remain poorly understood. More generally, it can be difficult to reason about the complex web of direct and indirect species interactions in real assemblages, especially when these interactions occur against a background of other ecological processes such as dispersal limitation 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, versus 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 generally provide convincing evidence of species' direct effects on one another.

While 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 (Lee and Hastie 2012, Loh and Wainwright 2013), have been used in many scientific fields for decades to model associations between various kinds of "particles". For example, a well-studied network named 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 the same 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

70 one; Gilpin and Diamond 1982). Now that it is computationally feasible to fit these models
71 exactly, the approach has become worth a second look.

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

78 *Methods*

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

90 Markov networks give a probability value for every possible combination of presences and
91 absences in communities. For example, given a network with binary outcomes (i.e. 0 for
92 absence and 1 for presence), the relative probability of observing a given presence-absence

4

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 an intercept term determining 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; it controls the probability that 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 is defined so that 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 spirals exponentially into the billions

and beyond. Numerous techniques are available for working with Markov networks that keep

the computations tractable, e.g. via 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 model) 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 replicate landscapes with

varying numbers of species and sites (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 biotic interactions 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. In the previous two

sections, the values of α and β were known. In practice, however, ecologists will often need

to estimate these parameters from co-occurrence data. 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

7

$_{159}$ and gradient as R code. The rosalia package then uses the `BFGS` method in R's `optim`

$_{160}$ function to find the best values for α and β.

$_{161}$ For analyses with 5 or more species, I made a small modification to the maximum likelihood

$_{162}$ procedure described above. Given the large number of parameters associated with some of

$_{163}$ the networks to be estimated, I regularized the likelihood using a logistic prior distribution

$_{164}$ (Gelman et al. 2008) with a scale of 1 on the α and β terms.

$_{165}$ ***Other inference techniques for comparison.*** After fitting Markov networks to the

$_{166}$ simulated landscapes described above, I used several other techniques for inferring the sign

$_{167}$ and strength of marginal associations between pairs of species (Appendix B).

$_{168}$ The first two alternative interaction measures were the sample covariances and the partial

$_{169}$ covariances between each pair of species' data vectors on the landscape (Albrecht and Gotelli

$_{170}$ 2001). Because partial covariances are undefined for landscapes with perfectly-correlated

$_{171}$ species pairs, I used a regularized estimate based on ridge regression [Wieringen and Peeters

$_{172}$ (2014); i.e. linear regression with a Gaussian prior]. For these analyses, I set the ridge

$_{173}$ parameter to 0.2 divided by the number of sites on the landscape.

$_{174}$ The third alternative method, described in Gotelli and Ulrich (2009), involved simulating

$_{175}$ possible landscapes from a null model that retains the row and column sums of the original

$_{176}$ matrix (Strong et al. 1984). Using the default options in the Pairs software described in

$_{177}$ Gotelli and Ulrich (2009), I simulated the null distribution of scaled C-scores (a test statistic

$_{178}$ describing the number of *non*-co-occurrences between two species). The software then

$_{179}$ calculated a Z statistic for each species pair using this null distribution. After multiplying

$_{180}$ this statistic by -1 so that positive values corresponded to facilitation and negative values

$_{181}$ corresponded to competition, I used it as another estimate of species interactions.

8

Method evaluation. For the simulated landscapes based on Figure 1, method evaluation was fairly qualitative: any method whose test statistic for the two shrubs indicated a negative relationship 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 method, 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. These regressions yielded squared errors for each of the 23,140 simulated species pairs, across all conditions and replicates. I then partitioned these errors into groups defined by the properties of the simulated landscapes (i.e. species richness, number of observed communities, and the presence/absence of environmental filtering). Within each partition, the mean squared error was used to calculate the proportion of variance explained (compared with a baseline model that assumed all interaction strengths to be zero).

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 for 94% of the simulated landscapes, 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,

9

depending on the parameters that produced the simulated communities (Figure 3). In general, however, there was a consistent ordering: overall, the Markov network explained 54% of the "true" parameters' squared deviations from zero, followed by partial covariances (33%), and sample covariances (22%). The null model scores initially explained only 12%. After manually reducing the value of one especially strong outlier ($Z = 1004$, implying $p < 10^{-1000000}$), this increased to 17% (Appendix B). Figure 3 reflects the adjusted version of the results.

Discussion

The results presented above show that Markov networks can reliably recover species' pairwise interactions from observational data, even for cases where a common null modeling technique reliably fails. Specifically, Markov networks were successful even when direct interactions were largely overwhelmed by indirect effects (Figure 1) or environmental effects (lower panels of Figure 3). For cases where fitting a Markov network is computationally infeasible, these results also indicate that partial covariances—which can be computed straightforwardly by linear regression—can often provide a surprisingly useful approximation. The partial correlations' success on simulated data may not carry over to real data sets, however; Loh and Wainwright (2013) show that the linear approximations can be less reliable in cases where the true interaction matrix contains more structure (e.g. guilds or trophic levels). On the other hand, if ecologists are familiar enough with the natural history of their study systems to describe this kind of structure as a prior distribution on the parameters or as a penalty on the likelihood, then this information could reduce the effective degres of freedom to estimate and real-world results might be even better than those shown in Figure 3.

Ecologists will also need natural history to pin down the exact nature of the interactions

identified by a network model (e.g. which species in a positively-associated pair is facilitating the other), particularly when real pairs of species can reciprocally influence one another in multiple ways simultaneously (Bruno et al. 2003); the β coefficients in Markov networks have to reduce this complexity to a single number. In short, partial correlations and Markov networks both help prevent us from mistaking marginal associations for conditional ones, but they can't tell us the underlying biological mechanisms at work.

Despite these limitations, Markov networks have enormous potential to improve ecological inferences. For example, Markov networks provide a simple answer to the question of how competition should affect a species' overall prevalence, which was a major flash point for the null model debates in the 1980's (Roughgarden 1983, Strong et al. 1984). Equation 1 can be used to calculate the expected prevalence of a species in the absence of biotic influences ($\frac{1}{1+e^{-\alpha}}$; Lee and Hastie 2012). Competition's effect on prevalence in a Markov network can then be calculated by subtracting this value and the observed prevalence (cf Figure 2D).

Markov networks—particularly the Ising model for binary networks—have been studied in statistical physics for nearly a century (Cipra 1987), and the models' properties, capabilities, and limits are well-understood in a huge range of applications, from spatial statistics (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 (i.e. phase transitions; Cipra (1987)).

This modeling approach is also 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

11

assemblages (>300 species) while simultaneously modeling each species' nonlinear response to the abiotic environment. Combining multiple ecological processes into a common model could help ecologists to disentangle different factors that can confound simpler co-occurrence analyses (cf Connor et al. 2013). Numerous other extensions are possible: Markov networks can be fit with a mix of discrete and continuous variables, for example (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 tip the balance between facilitation and exploitation (Bruno et al. 2003).

Finally, the results presented here have important implications for ecologists' continued use of null models to draw inferences about species interactions. 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 (Gotelli and Ulrich 2009). Estimating pairwise coefficients directly seems like a much more promising approach: to the extent that the models' relative performance on real data sets is similar to the range of results shown in Figure 3, scientists in this field could easily double their explanatory power by switching from null models to linear regression and partial covariances, or triple it by switching to a Markov network.

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.

13

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].

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

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

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

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

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

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

327 Roughgarden, J. 1983. Competition and Theory in Community Ecology. The American
328 Naturalist 122:583–601.

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

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

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

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

338 Whittam, T. S., and D. Siegel-Causey. 1981. Species Interactions and Community Structure

339 in Alaskan Seabird Colonies. Ecology 62:1515–1524.

340 Wieringen, W. N. van, and C. F. Peeters. 2014. Ridge Estimation of Inverse Covariance

341 Matrices from High-Dimensional Data. arXiv preprint arXiv:1403.0904.

342 Xiao, X., D. J. McGlinn, and E. P. White. 2015. A strong test of the Maximum Entropy

343 Theory of Ecology. The American Naturalist 185:E70–E80.

344 ***Figure captions***

345 ***Figure 1.*** **A.** A small network of three competing species. The tree (top) tends not to

346 co-occur with either of the two shrub species, as indicated by the strongly negative

347 coefficient linking them. The two shrub species also compete with one another, as indicated

348 by their negative coefficient (circled), but this effect is substantially weaker. **B.** In spite of

349 the competitive interactions between the two shrub species, their shared tendency to occur

350 in locations without trees makes their occurrence vectors positively correlated (circled). **C.**

351 Controlling for the tree species' presence with a conditional method such as a partial

352 covariance or a Markov network allows us to correctly identify the negative interaction

353 between these two species (circled).

354 ***Figure 2.*** **A.** A small Markov network with two species. The depicted abiotic environment

355 favors the occurrence of both species ($\alpha > 0$), particularly species 2 ($\alpha_2 > \alpha_1$). The negative

356 β coefficient linking these two species implies that they co-occur less than expected under

357 independence. **B.** Relative probabilities of all four possible presence-absence combinations

358 for Species 1 and Species 2. The exponent includes α_1 whenever Species 1 is present ($y_1 = 1$),

359 but not when it is absent ($y_1 = 0$). Similarly, the exponent includes α_2 only when species 2 is

360 present ($y_2 = 1$), and β only when both are present ($y_1 y_2 = 1$). The normalizing constant Z,

361 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.