1 **Title:** Inferring species interactions from co-occurrence data with Markov networks

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

3 **Abstract:** Inferring species interactions from co-occurrence data is one of the most

4 controversial tasks in community ecology. One difficulty is that a single pairwise interaction

5 can ripple through an ecological network and produce surprising indirect consequences. For

6 example, the negative correlation between two competing species can be reversed in the

7 presence of a third species that is capable of outcompeting both of them. Here, I apply

8 models from statistical physics, called Markov networks or Markov random fields, that can

9 predict the direct and indirect consequences of any possible species interaction matrix.

10 Interactions in these models can also be estimated from observed co-occurrence rates via

11 maximum likelihood, controlling for indirect effects. Using simulated landscapes with known

12 pairwise interaction strengths, I evaluated Markov networks and six existing approaches. The

13 Markov networks consistently outperformed other methods, correctly isolating direct

14 interactions between species pairs even when indirect interactions or abiotic factors largely

15 overpowered them. Two computationally efficient approximations, based on controlling for

16 indirect effects with linear or generalized linear models, also performed well. Indirect effects

17 reliably caused a common null modeling approach to produce incorrect inferences, however.

Introduction

21 To the extent that nontrophic species interactions (such as competition) affect community

22 assembly, ecologists might expect to find signatures of these interactions in species

23 composition data (MacArthur 1958, Diamond 1975). Despite decades of work and several

24 major controversies, however (Lewin 1983, Strong et al. 1984, Connor et al. 2013), existing

1

25 methods for detecting competition's effects on community structure are unreliable (Gotelli

26 and Ulrich 2009). In particular, species' effects on one another can become lost in a web of

27 indirect effects. For example, the competitive interaction between the two shrub species in

28 Figure 1A is obscured by their shared tendency to occur in unshaded areas (Figure 1B).

29 While ecologists have long known that indirect effects can overwhelm direct ones at the

30 landscape level (Dodson 1970, Levine 1976), the vast majority of our methods for drawing

31 inferences from observational data do not control for these effects (e.g. Diamond 1975, Strong

32 et al. 1984, Gotelli and Ulrich 2009, Veech 2013, Pollock et al. 2014). To the extent that

33 indirect interactions like those in Figure 1 are generally important (Dodson 1970), existing

34 methods will not provide much evidence regarding species interactions.

35   While competition doesn't reliably reduce co-occurrence rates at the whole-landscape

36 level (as most methods assume), it nevertheless does leave a signal in the data (Figure 1C).

37 For example, after partitioning the data set into shaded and unshaded sites, there will be

38 co-occurrence deficits in each subset that wouldn't otherwise be apparent. More generally,

39 controlling for other species in the network will often be important for obtaining reliable

40 estimates of direct (conditional, or all-else-equal) effects. This kind of precision is difficult to

41 obtain from null models, which only simulate the most extreme possible scenario: that *all*

42 direct and indirect interactions are exactly zero. Nevertheless, null models have dominated

43 this field for more than three decades (Strong et al. 1984, Gotelli and Ulrich 2009).

44   Following Azaele et al. (2010), this paper shows that Markov networks (undirected

45 graphical models also known as Markov random fields; Murphy 2012) can provide a

46 framework for understanding the landscape-level consequences of pairwise species

47 interactions, and for estimating them from observed presence-absence matrices. Markov

48 networks have been used in many scientific fields in similar contexts for decades, from physics

2

⁴⁹ (where nearby particles interact magnetically; Cipra 1987) to spatial statistics (where

⁵⁰ adjacent grid cells have correlated values; Harris 1974, Gelfand et al. 2005). While

⁵¹ community ecologists explored some related approaches in the 1980's (Whittam and

⁵² Siegel-Causey 1981), they used severe approximations that led to unintelligible results (e.g.

⁵³ "probabilities" greater than one; Gilpin and Diamond 1982).

⁵⁴ Below, I demonstrate Markov networks' ability to produce exact predictions about the

⁵⁵ direct and indirect consequences of an interaction matrix, and also to make inferences about

⁵⁶ the species interactions that contributed to an observed set of co-occurrences. Using

⁵⁷ simulated data sets where the "true" interactions are known, I compare this approach with

⁵⁸ several existing methods. Finally, I discuss opportunities for extending the approach

⁵⁹ presented here to other problems in community ecology, e.g. quantifying the overall effect of

⁶⁰ species interactions on occurrence rates (Roughgarden 1983) and disentangling the effects of

⁶¹ biotic versus abiotic interactions on species composition (Pollock et al. 2014).

⁶² **Methods**

⁶³ **Markov networks.** Markov networks provide a framework for translating back and forth

⁶⁴ between the conditional (all-else-equal) relationships among species (Figure 1C) and the

⁶⁵ kinds of species assemblages that these relationships produce. Here, I show how a set of

⁶⁶ conditional relationships can be used to determine how groups of species can co-occur.

⁶⁷ Methods for estimating conditional relationships from data are discussed in the next section.

⁶⁸ A Markov network defines the relative probability of observing a given vector of

⁶⁹ species-level presences (1s) and absences (0s), $\vec{y}$ at a site, as

⁷⁰
$$p(\vec{y}; \alpha, \beta) \propto exp(\sum_i \alpha_i y_i + \sum_{<ij>} \beta_{ij} y_i y_j),$$

⁷¹ where the second sum is over all $\frac{1}{2}n(n-1)$ pairs of n species. In this model, α_i is an

⁷² intercept term determining the amount that the presence of species i contributes to the

3

73 log-probability of $\vec{y}$; it directly controls the prevalence of species i. Similarly, β_{ij} is the

74 amount that the co-occurrence of species i and species j contributes to the log-probability; it

75 controls the conditional relationship between two species, i.e. the probability that they will

76 be found together, after controlling for the other species in the network (Figure 2A, Figure

77 2B). For example, if $\beta_{ij} = +2$, then each species' odds of occurrence would be e^2 times higher

78 when the other one is present (as compared with otherwise equivalent sites). Because the

79 relative probability of a presence-absence vector increases when positively-associated species

80 co-occur and decreases when negatively-associated species co-occur, the model tends—all else

81 equal—to produce assemblages that have many pairs of positively-associated species and

82 relatively few pairs of negatively-associated species (exactly as an ecologist might expect). Of

83 course, if all else is *not* equal (e.g. Figure 1, where the presence of one competitor is

84 associated with release from another competitor), then predicting species' overall

85 co-occurrence rates can be more complicated, and may require summing over the different

86 possible assemblages (Figure 2B).

87 **Estimating α and β coefficients from presence-absence data.** In the previous section,

88 the values of α and β were known and the goal was to make predictions about possible species

89 assemblages. In practice, however, ecologists will often need to estimate the parameters from

90 an observed co-occurrence matrix (i.e. from a set of independent $\vec{y}$ vectors indicating which

91 species are present at each site on the landscape). When the number of species is reasonably

92 small, one can compute exact maximum likelihood estimates for all of the α and β

93 coefficients given a presence-absence matrix by optimizing $p(\vec{y}; \alpha, \beta)$. Fully-observed Markov

94 networks like the ones considered here have unimodal likelihood surfaces (Murphy 2012),

95 ensuring that this procedure will always converge on the global maximum. This maximum

96 represents the unique combination of α and β coefficients that would be expected to produce

exactly the observed co-occurrence frequencies on average (i.e., maximizing the likelihood ensures that the sufficient statistics of the model distribution match the sufficient statistics of the data; Murphy 2012). I used the rosalia package (Harris 2015a) for the R programming language (R Core Team 2015) to optimize the Markov network parameters.

Simulated landscapes. I simulated several sets of landscapes using known parameters to evaluate different statistical methods' performance. The first set of landscapes included the three competing species shown in Figure 1. For each of 1000 replicates, I generated a landscape with 100 sites by sampling from a probability distribution defined by the figure's interaction coefficients (Appendix 1). Each of the methods described below was then evaluated on its ability to correctly infer that the two shrub species competed with one another, despite their frequent co-occurrence.

I then generated landscapes with up to 20 interacting species at 25, 200, or 1600 sites using three increasingly complex models (50 replicates for each combination of size and model; see Appendix 2 for details). I randomly drew the "true" coefficient magnitudes for each replicate landscape from exponential distributions so that most species pairs interacted negligibly but a few pairs interacted strongly enough that their effects could propagate indirectly to other species in the network.

The first set of 20-species landscapes, like the landscapes with three species, were generated directly from a Markov network to ensure that the model could recover the parameters used to generate the "observed" co-occurrence data. Then, I added two environmental factors that varied from location to location across the simulated landscapes, and simulated a new set of co-occurrence data so that species' α coefficients depended on the local environment. The latter set of simulated landscapes provide an important test of the methods' ability to distinguish co-occurrence patterns that were generated from pairwise

₁₂₁ biotic interactions from those that were generated by external forces like abiotic

₁₂₂ environmental filtering. This task was made especially difficult because—as with most

₁₂₃ analyses of presence-absence data for co-occurrence patterns—the inference procedure did not

₁₂₄ have access to any information about the environmental or spatial variables that helped

₁₂₅ shape the landscape (cf Connor et al. 2013). I generated the final set of landscapes with an

₁₂₆ abundance-based model that included per-capita interaction rates instead of per-species

₁₂₇ interaction rates.

₁₂₈ **Recovering species interactions from simulated data.** I compared seven techniques

₁₂₉ for determining the sign and strength of the associations between pairs of species from

₁₃₀ simulated data (Appendix 4). First, I used the rosalia package (Harris 2015a) to fit Markov

₁₃₁ network models, as described above. For the analyses with 20 species, I added a weak

₁₃₂ regularizer (equivalent to a logistic prior with scale 2) to ensure that the model estimates

₁₃₃ were always finite.

₁₃₄ I also evaluated six alternative methods: five from the existing literature, plus a novel

₁₃₅ combination of two of these methods. The first alternative interaction metric was the sample

₁₃₆ correlation between species' presence-absence vectors, which summarizes their marginal

₁₃₇ association. Next, I used partial correlations, which summarize species' conditional

₁₃₈ relationships. This approach is common in molecular biology (Friedman et al. 2008), but is

₁₃₉ rare in ecology (see Albrecht and Gotelli (2001) and Faisal et al. (2010) for two exceptions).

₁₄₀ In the context of non-Gaussian data, the partial correlation can be thought of as a

₁₄₁ computationally efficient approximation to the full Markov network model (Loh and

₁₄₂ Wainwright 2013). Because partial correlations are undefined for landscapes with

₁₄₃ perfectly-correlated species pairs, I used a regularized estimate based on James-Stein

₁₄₄ shrinkage, as implemented in the corpcor package's `pcor.shrink` function with the default

145 settings (Schäfer et al. 2014).

146 The third alternative, generalized linear models (GLMs), also provide a computationally

147 efficient approximation to the Markov network (Lee and Hastie 2012). Following Faisal et al.

148 (2010), I fit regularized logistic regression models (Gelman et al. 2008) for each species, using

149 the other species on the landscape as predictors. This produced two interaction estimates for

150 each species pair; one for the effect of species i on species j and one for the reverse. These two

151 coefficients are not identifiable from the data, however (Schmidt and Murphy 2012), so I used

152 their average as an overall measure of the overall relationship between species i and species j.

153 The next method, described in Gotelli and Ulrich (2009), involved simulating new

154 landscapes from a null model that retains the row and column sums of the original matrix

155 (Strong et al. 1984). I used the Z-scores computed by the Pairs software described in Gotelli

156 and Ulrich (2009) as my null model-based estimator of species interactions.

157 The last two estimators used the latent correlation matrix estimated by the BayesComm

158 package (Golding and Harris 2015) in order to evaluate the recent claim that the correlation

159 coefficients estimated by "joint species distribution models" provide an accurate assessment

160 of species' pairwise interactions (Pollock et al. 2014, see also Harris 2015b). In addition to

161 using the posterior mean correlation (Pollock et al. 2014), I also used the posterior mean

162 *partial* correlation, which should control better for indirect effects.

163 **Evaluating model performance.** For the simulated landscapes based on Figure 1, I

164 assessed whether each method's test statistic indicated a positive or negative relationship

165 between the two shrubs (Appendix 1). For the null model (Pairs), I calculated statistical

166 significance using its Z-score. For the Markov network, I used the Hessian matrix to generate

167 approximate confidence intervals.

168 I then evaluated the relationship between each method's estimates and the "true"

interaction strengths among all of the species pairs from the larger simulated landscapes. This determined which of the methods provide a consistent way to know how strong species interactions are—regardless of which species were present in a particular data set or how many observations were taken. To ensure that the different test statistics (e.g. correlations versus Z scores) were on a common scale, I rescaled them using linear regression through the origin. I then calculated the proportion of variance explained for different combinations of model type and landscape size (compared with a null 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 correlation between these species was positive in 94% of replicates, and the randomization-based null model falsely reported positive associations 100% of the time. Worse, more than 98% of these false conclusions were statistically significant. The partial correlation and Markov network estimates, on the other hand, each correctly isolated the direct negative interaction between the shrubs from their positive indirect interaction 94% of the time (although the confidence intervals overlapped zero in most replicates).

Twenty species. In general, each model's performance was highest for large landscapes with simple assembly rules and no environmental heterogeneity (Figure 3). Despite some variability across contexts, the rank ordering across methods was very consistent. In particular, the four methods that controlled for indirect effects (the Markov network, the generalized linear models, and the two partial correlation-based methods) always matched or outperformed those that did not.

193   Of these, the Markov network consistently performed the best. As anticipated by Lee

194 and Hastie (2012), generalized linear models closely approximated the Markov network

195 estimates, especially when the data sets were very large [[Figure]]. As reviewed in Gotelli and

196 Ulrich (2009), however, most analyses in this field of ecology involve fewer than 50 sites, so

197 model performance on the smaller landscapes (where the gap was larger) is very important.

198 As shown in Appendix 5, the standard errors associated with the estimates in Figure 3 are

199 small (less than 0.01), so the differences among methods should not be attributed to

200 sampling error.

201   Of the methods that did not control for indirect effects, Figure 3 shows that simple

202 correlation coefficients provided a more reliable indicator of species' true interaction strengths

203 than either the joint species distribution model (BayesComm) or the null model (Pairs). The

204 estimates from these approaches were tightly correlated (after controlling for the size of the

205 landscape) suggesting that the null model only contains a noisy version of the same

206 information that could be obtained more easily and interpretably with simple correlation

207 coefficients (Figure 4B).

208   Finally, we can evaluate the models' statistical inferences (focusing on the first two

209 simulation types, for which the true interaction rates are easiest to interpret). The Markov

210 network's approximate Type I error rate (defined here as the probability that 0 fell outside

211 the 95% confidence interval for a pair of species where $|\beta_{ij}| < 0.1$) depended on the

212 simulation type: 0.02 for simulations that matched the model's assumptions, versus 0.14 for

213 simulations that included environmental heterogeneity (see Appendix 5 for confidence

214 interval coverage across a range of β_{ij} values). The null model's Type I error rates were

215 [[XX]] and [[XX]] for the constant and heterogeneous landscape, respectively—far higher

216 than the nominal 0.05 rate. Figure 4C shows, across a range of true interaction strengths, the

9

²¹⁷ probability that the null model or the Markov network will predict the wrong sign of the

²¹⁸ interaction with 95% confidence. The null model makes such errors [[XX]] times as often as

²¹⁹ the Markov network, even though it only reject the null hypothesis [[XX]] as often overall

²²⁰ (Appendix 5). The Markov network's errors were also more concentrated around 0, so it

²²¹ never misclassified strong interactions like the null model did.

²²² **Discussion**

²²³ The results presented above show that Markov networks can reliably recover species' pairwise

²²⁴ interactions from species composition data, even for cases where environmental heterogeneity

²²⁵ and indirect interactions cause ecologists' typical null modeling approaches to reliably fail.

²²⁶ Partial covariances and generalized linear models can both provide computationally efficient

²²⁷ approximations, but with somewhat lower accuracy, especially for typically-sized data sets

²²⁸ with small numbers of sites (Gotelli and Ulrich 2009). The difference in accuracy may be

²²⁹ larger for real data sets than for the simulated landscapes in Figure 3, however; linear

²³⁰ approximations to the Markov network make larger errors when the interaction matrix is

²³¹ structured (e.g. due to guilds or trophic levels; Loh and Wainwright 2013). Similarly, the

²³² separate generalized linear models for each species can severely overfit in some cases (Lee and

²³³ Hastie 2012). The full Markov network should thus be preferred to the approximations when

²³⁴ it is computationally tractable.

²³⁵ Compositional data only contains enough degrees of freedom to estimate one interaction

²³⁶ per species pair (Schmidt and Murphy 2012), so none of these methods can identify the exact

²³⁷ nature of the pairwise interactions (e.g. which species in a positively-associated pair is

²³⁸ facilitating the other). To estimate asymmetric interactions, such as commensalism or

²³⁹ predation, ecologists could use time series, behavioral observations, manipulative

²⁴⁰ experiments, or natural history. These other sources of information could also be used to

10

241 augment the likelihood function with an informative prior distribution, reducing ecologists'

242 error and uncertainty relative to Figure 3's results.

243 Despite their limitations, Markov networks have enormous potential to improve

244 ecological understanding. In particular, they make many fewer errors than existing

245 approaches, and can make precise statements about the conditions where indirect

246 interactions will overwhelm direct ones. They also provide a simple answer to the question of

247 how competition should affect a species' overall prevalence, which has important implications

248 for community-level modeling (Strong et al. 1984). Specifically, Equation 1 can be used to

249 calculate the expected prevalence of a species in the absence of biotic influences as

250 $e^{\alpha}/(e^0 + e^{\alpha})$. Competition's effect on prevalence can then be estimated by comparing this

251 value with the observed prevalence (e.g. comparing Figure 2D with Figure 2C). This novel

252 quantitative result conflicts with most of our null models, which unreasonably assume that

253 prevalence would be the exactly same in the absence of competition as it is in the observed

254 data (Roughgarden 1983).

255 Markov networks—particularly the Ising model for binary networks—are very well

256 understood, having been studied for nearly a century (Cipra 1987). Tapping into this

257 framework would thus allow ecologists to take advantage of into a vast set of existing

258 discoveries and techniques for dealing with indirect effects, stability, and alternative stable

259 states. Numerous extensions to the basic network are possible as well. For example, the

260 states of the interaction network can be modeled as a function of the local abiotic

261 environment (Lee and Hastie 2012), which would help incorporate networks of biotic

262 interactions into species distribution models (Pollock et al. 2014) and lead to a better

263 understanding of the interplay between biotic and abiotic effects on community structure.

264 Alternatively, models could allow one species to alter the relationship between two other

11

265 species (Tjelmeland and Besag 1998, cf Bruno et al. 2003).

266 Finally, the results presented here have important implications for ecologists' continued

267 use of null models for studying species interactions. Null and neutral models can be useful for

268 clarifying our thinking (Harris et al. 2011, Xiao et al. 2015), but deviations from a particular

269 null model must be interpreted with care (Roughgarden 1983). Even in small networks with

270 three species, it may simply not be possible to implicate specific ecological processes like

271 competition by rejecting a general-purpose null (Gotelli and Ulrich 2009), especially when the

272 test statistic is effectively just a correlation coefficient (Figure 4B). When the non-null

273 backdrop is not controlled for, Type I error rates can skyrocket, the apparent sign of the

274 interaction can change, and null models can routinely produce misleading inferences (Figure

275 1, Figure 4C, Gotelli and Ulrich (2009)).

276 Controlling for indirect effects via simultaneous estimation of multiple ecological

277 parameters seems like a much more promising approach: to the extent that the models'

278 relative performance on real data sets is similar to the range of results shown in Figure 3,

279 scientists in this field could more than double their explanatory power by switching from null

280 models to Markov networks (or increase it nearly as much with linear or generalized linear

281 approximations). Regardless of the methods ecologists ultimately choose, controlling for

282 indirect effects could clearly improve our understanding of species' direct effects on one

283 another and on community structure.

Science Foundation and by the Gordon and Betty Moore Foundation's Data-Driven

Discovery Initiative through Grant GBMF4563 to E. P. White.

References:

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, but more weakly

(circled coefficient). **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 trees with a conditional (all-else-equal)

approach such as a partial covariance or a Markov network leads to correct identification of

the negative shrub-shrub interaction (circled).

Figure 2: A. A small Markov network, defined by its α and β values. The abiotic

environment favors the occurrence of each species ($\alpha > 0$), particularly species 2 ($\alpha_2 > \alpha_1$).

The negative β_{12} coefficient is consistent with competition between the two species. **B.** The

coefficients determine the probabilities of all four possible presence-absence combinations for

Species 1 and Species 2. α_1 is added to the exponent 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 includes β_{12} only when both are present ($y_1 y_2 = 1$). The normalizing

constant Z, ensures that the four probabilities sum to 1. In this case, Z is about 18.5. **C.**

The expected frequencies of all possible co-occurrence patterns between the two species of

interest, as calculated in the previous panel. **D.** Without competition (i.e. with $\beta_{12} = 0$, each

species would occur more often.

Figure 3: Proportion of variance in interaction coefficients explained by each method versus

number of sampled locations across the three simulation types.

13

Figure 4: A. The Markov network's estimated interaction coefficients were generally very similar to the GLM estimates. **B.** The null model's estimates typically matched the (negative) correlation coefficient, after controlling for landscape size. **C.** For any given interaction strength, the null model was much more likely to misclassify its sign with 95% confidence than the Markov network was.

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.

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.

Dodson, S. I. 1970. COMPLEMENTARY FEEDING NICHES SUSTAINED BY SIZE-SELECTIVE PREDATION. Limnology and Oceanography 15:131–137.

Faisal, A., F. Dondelinger, D. Husmeier, and C. M. Beale. 2010. Inferring species interaction networks from species abundance data: A comparative evaluation of various statistical and machine learning methods. Ecological Informatics 5:451–464.

Friedman, J., T. Hastie, and R. Tibshirani. 2008. Sparse inverse covariance estimation with

336   the graphical lasso. Biostatistics 9:432–441.

337   Gelfand, A. E., A. M. Schmidt, S. Wu, J. A. Silander, A. Latimer, and A. G. Rebelo. 2005.

338   Modelling species diversity through species level hierarchical modelling. Journal of the Royal

339   Statistical Society: Series C (Applied Statistics) 54:1–20.

340   Gelman, A., A. Jakulin, M. G. Pittau, and Y.-S. Su. 2008. A Weakly Informative Default

341   Prior Distribution for Logistic and Other Regression Models. The Annals of Applied

342   Statistics 2:1360–1383.

343   Gilpin, M. E., and J. M. Diamond. 1982. Factors contributing to non-randomness in species

344   Co-occurrences on Islands. Oecologia 52:75–84.

345   Golding, N., and D. J. Harris. 2015. BayesComm: Bayesian Community Ecology Analysis.

346   Gotelli, N. J., and W. Ulrich. 2009. The empirical Bayes approach as a tool to identify

347   non-random species associations. Oecologia 162:463–477.

348   Harris, D. J. 2015a. Rosalia: Exact inference for small binary Markov networks. R package

349   version 0.1.0. Zenodo. http://dx.doi.org/10.5281/zenodo.17808.

350   Harris, D. J. 2015b. Generating realistic assemblages with a Joint Species Distribution Model.

351   Methods in Ecology and Evolution.

352   Harris, D. J., K. G. Smith, and P. J. Hanly. 2011. Occupancy is nine-tenths of the law:

353   Occupancy rates determine the homogenizing and differentiating effects of exotic species.

354   The American naturalist 177:535.

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

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

357   Levine, S. H. 1976. Competitive Interactions in Ecosystems. The American Naturalist

358   110:903–910.

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

360 Loh, P.-L., and M. J. Wainwright. 2013. Structure estimation for discrete graphical models:

361 Generalized covariance matrices and their inverses. The Annals of Statistics 41:3022–3049.

362 MacArthur, R. H. 1958. Population ecology of some warblers of northeastern coniferous

363 forests. Ecology 39:599–619.

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

365 Pollock, L. J., R. Tingley, W. K. Morris, N. Golding, R. B. O'Hara, K. M. Parris, P. A. Vesk,

366 and M. A. McCarthy. 2014. Understanding co-occurrence by modelling species

367 simultaneously with a Joint Species Distribution Model (JSDM). Methods in Ecology and

368 Evolution:n/a–n/a.

369 R Core Team. 2015. R: A Language and Environment for Statistical Computing. R

370 Foundation for Statistical Computing, Vienna, Austria.

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

372 Naturalist 122:583–601.

373 Schäfer, J., R. Opgen-Rhein, V. Zuber, M. Ahdesmäki, A. P. D. Silva, and K. Strimmer.

374 2014. Corpcor: Efficient Estimation of Covariance and (Partial) Correlation.

375 Schmidt, M., and K. Murphy. 2012. Modeling Discrete Interventional Data using Directed

376 Cyclic Graphical Models. arXiv preprint arXiv:1205.2617.

377 Strong, D. R., D. Simberloff, L. G. Abele, and A. B. Thistle. 1984. Ecological communities:

378 Conceptual issues and the evidence. Princeton University Press.

379 Tjelmeland, H., and J. Besag. 1998. Markov Random Fields with Higher-order Interactions.

380 Scandinavian Journal of Statistics 25:415–433.

381 Veech, J. A. 2013. A probabilistic model for analysing species co-occurrence. Global Ecology

16

382   and Biogeography 22:252–260.

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

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

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

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