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

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

Abstract: Inferring species interactions from co-occurrence 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, the negative correlation between two competing species can be reversed in the presence of a third species that is capable of outcompeting both of them. 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 also be estimated from observed co-occurrence rates via maximum likelihood, controlling for indirect effects. Using simulated landscapes with known pairwise interaction strengths, I evaluated Markov networks and six existing approaches. The Markov networks consistently outperformed other methods, correctly isolating direct interactions between species pairs even when indirect interactions or abiotic factors largely overpowered them. Two computationally efficient approximations, based on controlling for indirect effects with linear or generalized linear models, also performed well. Indirect effects reliably caused a common null modeling approach to produce incorrect inferences, however.

Introduction

To the extent that nontrophic species interactions (such as competition) affect community assembly, ecologists might expect to find signatures of these interactions in species composition data (MacArthur 1958, Diamond 1975). Despite decades of work and several 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

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), $\vec{y}$, 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

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

4

97 exactly the observed co-occurrence frequencies on average (i.e., maximizing the likelihood

98 ensures that the sufficient statistics of the model distribution match the sufficient statistics of

99 the data; Murphy 2012). I used the rosalia package (Harris 2015a) for the R programming

100 language (R Core Team 2015) to optimize the Markov network parameters.

101 **Simulated landscapes.** I simulated several sets of landscapes using known parameters to

102 evaluate different statistical methods' performance. The first set of landscapes included the

103 three competing species shown in Figure 1. For each of 1000 replicates, I generated a

104 landscape with 100 sites by sampling from a probability distribution defined by the figure's

105 interaction coefficients (Appendix 1). Each of the methods described below was then

106 evaluated on its ability to correctly infer that the two shrub species competed with one

107 another, despite their frequent co-occurrence.

108 I then generated landscapes with up to 20 interacting species at 25, 200, or 1600 sites

109 using three increasingly complex models (50 replicates for each combination of size and

110 model). As described in Appendix [[X]], I randomly drew the "true" coefficient values for

111 each replicate landscape so that most species pairs interacted negligibly, a few pairs

112 interacted very strongly, and competition was three times more common than facilitation.

113 The first set of 20-species landscapes, like the landscapes with three species, were

114 generated directly from a Markov network to ensure that the model could recover the

115 parameters used to generate the "observed" co-occurrence data. Then, I added two

116 environmental factors that varied from location to location across the simulated landscapes,

117 and simulated a new set of co-occurrence data so that species' α coefficients depended on the

118 local environment. The latter set of simulated landscapes provide an important test of the

119 methods' ability to distinguish co-occurrence patterns that were generated from pairwise

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

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

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

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

124 shape the landscape (cf Connor et al. 2013). I generated the final set of landscapes by

125 simulating the population dynamics of 20 interacting species, as described in Appendix [[X]].

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

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

128 simulated data (Appendix 3). First, I used the rosalia package (Harris 2015a) to fit Markov

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

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

131 were always finite [[Appendix]].

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

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

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

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

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

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

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

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

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

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

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

143 settings (Schäfer et al. 2014).

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

6

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

partial correlation, which might be able to control for indirect effects.

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

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

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

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

approximate confidence intervals.

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

169 interactions are—regardless of which species were present in a particular data set or how

170 many observations were taken. To ensure that the different test statistics (e.g. correlations

171 versus Z scores) were on a common scale, I rescaled them using linear regression through the

172 origin. I then calculated the proportion of variance explained for different combinations of

173 model type and landscape size (compared with a null model that assumed all interaction

174 strengths to be zero).

Results

176 **Three species.** As shown in Figure 1, the marginal relationship between the two shrub

177 species was positive—despite their competition for space at a mechanistic level—due to

178 indirect effects of the dominant tree species. As a result, the correlation between these

179 species was positive in 94% of replicates, and the randomization-based null model falsely

180 reported positive associations 100% of the time. Worse, more than 98% of these false

181 conclusions were statistically significant. The partial correlation and Markov network

182 estimates, on the other hand, each correctly isolated the direct negative interaction between

183 the shrubs from their positive indirect interaction 94% of the time (although the confidence

184 intervals overlapped zero in most replicates).

185 **Twenty species.** Despite some variability across contexts (Figure 3), the four methods that

186 controlled for indirect effects (the Markov network, the generalized linear models, and the

187 two partial correlation-based methods) outperformed those that did not. Of these, the

188 Markov network consistently performed the best. As anticipated by Lee and Hastie (2012),

189 generalized linear models closely approximated the Markov network estimates, especially

190 when the data sets were very large. As reviewed in Gotelli and Ulrich (2009), however, most

191 analyses in this field involve fewer than 50 sites, so model performance on the smaller

192 landscapes where the gap was larger is most important. As shown in Appendix [[X]], the

193 standard errors associated with the estimates in Figure 3 are small (less than 0.01), so the

194 differences among methods should not be attributed to sampling error.

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

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

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

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

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

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

201 coefficients (Figure 4B).

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

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

204 network's approximate Type I error rate (defined here as its probability of rejecting the null

205 hypothesis for a pair of species where $|\beta_{ij}| < 0.1$) depended on the simulation type: 0.03 for

206 simulations that matched the model's assumptions, versus 0.15 for simulations that included

207 environmental heterogeneity [[Appendix]]. For the null model, these rates were 0.31 and 0.50,

208 respectively—far higher than the nominal 0.05 rate. Figure 4C shows, across a range of true

209 interaction strengths, the probability that the null model or the Markov network will predict

210 the wrong sign of the interaction with 95% confidence. The null model makes such errors

211 nearly 8 times as often as the Markov network, even though it only reject the null hypothesis

212 twice as often overall [[Appendix]]. The Markov network's errors were also more concentrated

213 around 0, so it never misclassified strong interactions like the null model did.

214 **Discussion**

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

216 interactions from observational data, even for cases where environmental heterogeneity and

9

217 indirect interactions cause ecologists' typical null modeling approaches to reliably fail. Partial

218 covariances and generalized linear models can both provide computationally efficient

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

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

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

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

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

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

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

226 it is computationally tractable.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

10

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

242 $e^\alpha/(e^0 + e^\alpha)$. Competition's effect on prevalence can then be estimated by comparing this

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

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

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

246 data (Roughgarden 1983).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

267 1, Figure 4C, Gotelli and Ulrich (2009)).

268   Controlling for indirect effects via simultaneous estimation of multiple ecological

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

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

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

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

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

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

275 another and on community structure.

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

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

285 linking them. The two shrub species also compete with one another, but more weakly

286 (circled coefficient). **B.** In spite of the competitive interactions between the two shrub species,

287 their shared tendency to occur in locations without trees makes their occurrence vectors

288 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.

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.

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.

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 the graphical lasso. Biostatistics 9:432–441.

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.

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

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

Levine, S. H. 1976. Competitive Interactions in Ecosystems. The American Naturalist 110:903–910.

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.

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

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

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

361 Evolution:n/a–n/a.

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

363 Foundation for Statistical Computing, Vienna, Austria.

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

365 Naturalist 122:583–601.

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

367 2014. Corpcor: Efficient Estimation of Covariance and (Partial) Correlation.

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

369 Cyclic Graphical Models. arXiv preprint arXiv:1205.2617.

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

371 Conceptual issues and the evidence. Princeton University Press.

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

373 Scandinavian Journal of Statistics 25:415–433.

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

375 and Biogeography 22:252–260.

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

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

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

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