Multi-Process Statistical Modeling of Species' Joint Distributions

By

DAVID JAY HARRIS
Bachelor of Arts (Washington University in St. Louis) 2008

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Population Biology

in the

OFFICE OF GRADUATE studies

of the

UNIVERSITY OF California

Davis

Approved:

asdf

df

Contents

1 Generating realistic assemblages with a Joint Species Distribution Model

David J. Harris

1.1 Introduction

A major goal of community ecology is to understand the processes, such as environmental filtering and species interactions, that determine where species could occur and which species can occur together (Chase 2003). Traditional multivariate methods for studying these issues in community ecology—such as ordination techniques for summarizing a data matrix's multivariate geometry—will not always provide the best approach to these questions, as they typically do not specify a data-generating mechanism or make predictions about new assemblages (but see Walker and Jackson 2011). More recent approaches, such as generalized linear models (Jackson et al. 2012, Wang et al. 2012, Jamil and Braak 2013) and species distribution models (SDMs; Elith and Leathwick 2009), can make specific predictions. Just as importantly, these predictions can be evaluated quantitatively based on their likelihoods.

Modern SDMs need not assume that species respond to environmental variation in a prespecified way (e.g. linearly or quadratically); relaxing this assumption has improved our ability to make predictions about individual species (Elith et al. 2006). For many community-level questions, however, species-level predictions may be of limited use. While SDMs can be combined ("stacked") to generate assemblage-level predictions (Pellissier et al. 2013), doing so implies that species' occurrence probabilities are uncorrelated (Clark et al. 2013, Calabrese et al. 2014). Ignoring the (potentially unobserved) factors driving these correlations can lead stacked models to generate incoherent jumbles of species rather than realistic assemblages (Clark et al. 2013). Given that most models only use climate variables as predictors (Austin and Van Niel 2011), the set of unobserved factors will usually include *all of ecology* apart from climatic influences. SDMs' failure to include other ecological processes is thus widely considered to be a major omission from statistical ecology's toolbox (Austin and Van Niel 2011, Guisan and Rahbek 2011, Kissling et al. 2012, Wisz et al. 2013, Clark et al. 2013).

In the last few years, several mixed models have been proposed to help explain the co-occurrence patterns that stacked SDMs ignore (Latimer et al. 2009, Ovaskainen et al. 2010, Golding 2013, Clark et al. 2013, Pollock et al. 2014). These *joint* species distribution models (JSDMs) can produce mixtures of possible species assemblages (points in Figure 1a), rather than relying on a small number of environmental measurements to fully describe each species' probability of occurrence (which would collapse the distribution in Figure 1a to a single point). In JSDMs (as in nature), a given set of climate estimates could be consistent with a number of different sets of co-occurring species, depending on factors that ecologists have not necessarily measured or even identified as important. JSDMs represent these unobserved (latent) factors as random variables whose true values are unknown, but whose existence would still help explain discrepancies between the data and the stacked SDMs' predictions (Figure 1). While JSDMs represent a major advance in community-level modeling (Clark et al. 2013, Pollock et al. 2014), existing implementations have all assumed that species' responses to the environment are linear (in the sense of a generalized linear model), limiting their accuracy and utility.

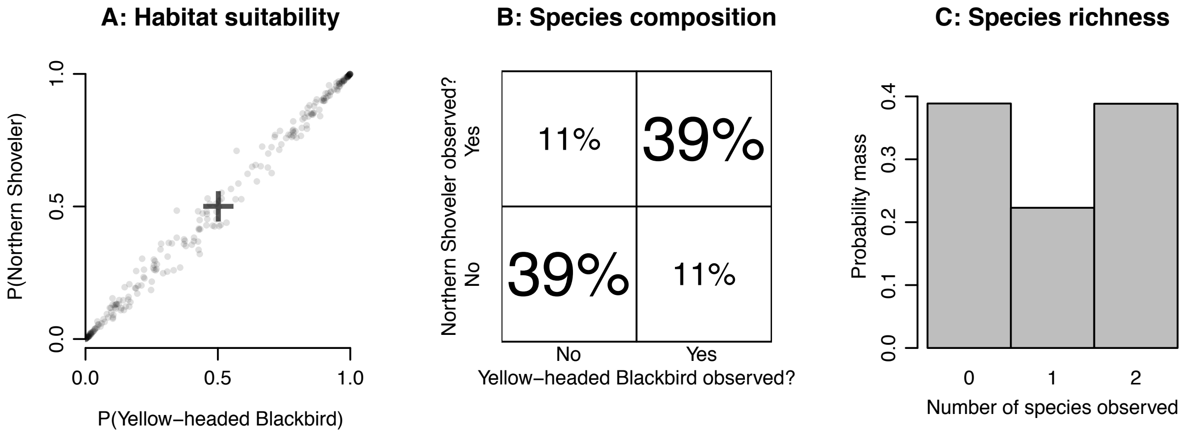


Figure 1.1: Unobserved environmental heterogeneity can induce correlations between species; ignoring this heterogeneity can produce misleading results. **A**: Based on climate predictors, a pair of single-species models might predict 50% occurrence probabilities for each of two wetland species (black cross). Climate predictors are not sufficient in this case, however: a site's suitability for these species cannot be fully determined without information about the availability of wetland habitat. Real habitats will to be tend to be suitable for both species (dense cloud of points in upper-right corner) or neither (lower-left corner), depending on this unmeasured variable. **B** This correlation among species substantially alters the set of assemblages one would expect to observe. (Under independence, all four possibilities would be equally probable.) **C** Positive correlations among species can even induce a strongly bimodal distribution of species richness values.

Here, I present a new R package for assemblage-level modeling—called *mistnet*—that does not rely on independence (as stacks of single-species models do) or linearity (as previous JSDMs have). Mistnet models are stochastic feed-forward neural networks (Neal 1992, Tang and Salakhutdinov 2013) that combine the flexibility of modern nonlinear models with the latent variables found in previous JSDMs. To demonstrate the value of this approach, I compared mistnet's predictive likelihood with that of several existing models, using observational data from thousands of North American Breeding Bird Survey transects (BBS; Sauer et al. 2011). A high predictive likelihood indicates that the model correctly expects to see the kinds of assemblages that were actually found out-of-sample, while a very low likelihood means that the model has effectively ruled those assemblages out due to overfitting or underfitting.

An accurate JSDM would up new possibilities for research and effective management. For example, although most models only have access to climate data (Austin and Van Niel 2011), a successful model of community structure should also be able to identify the major axes of non-climate variation that drive species turnover based on the species' observed co-occurrence patterns. Moreover, a successful assemblage-level model would be able to take advantage of the presence (or absence) of indicator species to inform its predictions about the rest of the assemblage. This ability to transfer information from easily-detected, well-documented taxa to more cryptic or rare species would prove valuable for community ecologists and conservationists alike.

1.2 Materials and Methods

Methods are presented in five main sections:

- A description of the data sets used
- An introduction to stochastic neural networks and the mistnet package
- The specific mistnet model used here
- A summary of the existing methods used for model comparison
- Criteria for model evaluation

1.2.1 Data

Field observations were obtained from the 2011 Breeding Bird Survey (BBS; Sauer et al. 2011). The BBS data consists of thousands of transects (“routes”), which served as the main unit for the analysis. Each route includes 50 stops, about 0.8 km apart. At each stop, all the birds observed in a 3-minute period are recorded, using a standardized procedure. Following BBS recommendations, I omitted nonstandard routes and data collected on days with bad weather.

To evaluate SDMs’ predictive capabilities, I split the routes into a “training” data set consisting of 1559 routes and a “test” data set consisting of 280 routes (Figure 2; Appendix A). The two data sets were separated by a 150-km buffer to ensure that models could not rely on spatial autocorrelation to make accurate predictions about the test set [c.f. Bahn and McGill (2007); Appendix A]. Each model was fit to the same training set, and then its performance was evaluated out-of-sample on the test set.

Observational data for each species was reduced to “presence” or “absence” at the route level, ignoring the possibility of observation error for the purposes of this analysis. 368 species were chosen for analysis according to a procedure described in Appendix A.

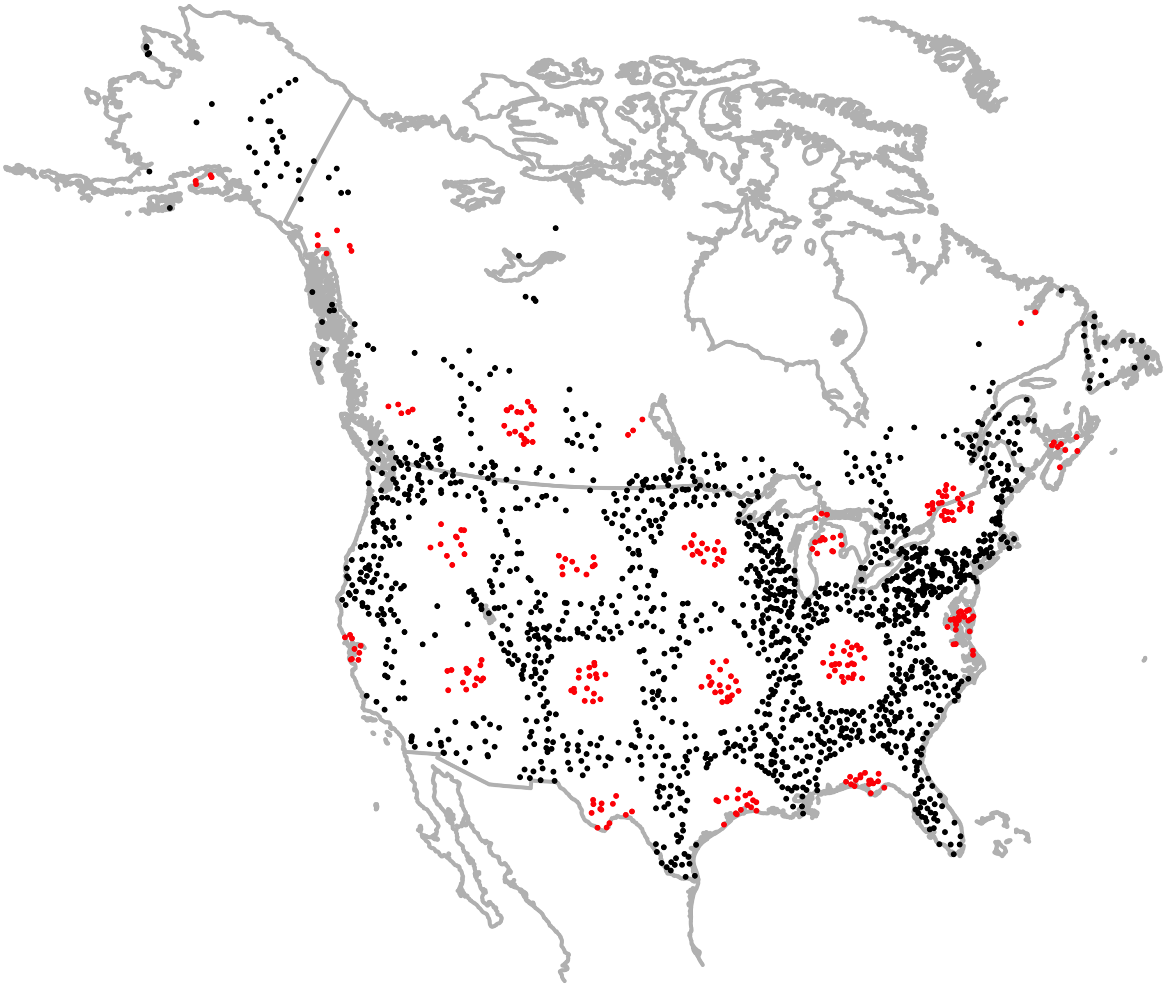


Figure 1.2: Map of the BBS routes used in this analysis. Black points are training routes; red ones are test routes. The training and test routes are separated by a 150-km buffer in order to minimize spatial autocorrelation across the two partitions.

I extracted the 19 Bioclim climate variables for each route from Worldclim (version 1.4; Hijmans et al. 2005) for use as environmental predictors. After removing predictors that were nearly collinear, eight climate-based predictors remained for the analyses (Appendix A). Since most SDMs do not use land cover data (Austin and Van Niel 2011) and one of mistnet's goals is to make inferences about unobserved environmental variation, no other variables were included in this analysis.

Finally, I obtained habitat classifications for each species from the Cornell Lab of Ornithology's All About Birds website (AAB; www.allaboutbirds.org) using an R script written by K. E. Dybala.

1.2.2 Introduction to stochastic neural networks

This section discusses the stochastic networks in general terms; the specific model used for avian communities is discussed in the following section. In general, ecologists have not had much success using neural networks for SDM (e.g. Dormann et al. 2008), but neural networks' recent success in other machine learning contexts (including contexts with latent random variables; Murphy 2012, Bengio 2013) makes them worth a second look for JSDM. While one can build stochastic versions of other nonlinear regression methods as well (e.g. Hutchinson et al. 2011), the relative simplicity of the backpropagation algorithm for training neural networks (Murphy 2012) makes them very appealing for exploratory research.

A neural net is a statistical model that makes its predictions by applying a series of nonlinear transformations to one or more predictor variables such as environmental measurements (Figure 3; Appendix B). After a suitable transformation of the environmental data, a final operation performs logistic regressions in the transformed space to make predictions about each species' occurrence probability (cf Leathwick et al. 2005). Training a neural network entails simultaneously optimizing the parameters associated with these transformations to optimize the overall likelihood (Appendix C).

Most neural networks' predictions are deterministic functions of their inputs. Applied to SDM, this would mean that each species' occurrence probability would be fully specified by the small number of variables that ecologists happen to measure. Mistnet's neural networks, in contrast, are *stochastic* [Neal (1992); Tang and Salakhutdinov (2013); Appendix B], meaning that they allow species' occurrence probabilities to depend on unobserved environmental factors as well. The true values of these unobserved factors are (by definition) not known, but one can still represent their *possible* values using samples from a probability distribution. In the absence of any information about what these variables should represent, mistnet defaults to sampling them from standard normal distributions. Depending on which values are sampled (i.e. on the possible states of the environment), the model could expect to see radically different kinds of species assemblages (Figure 1, Figure 3).

Inference can also proceed backward through a stochastic network: the presence (or absence) of one species provides information about the local environment, which can then be used to make better predictions about other species. For example, suppose that a researcher has more data about the local distribution of waterfowl—which are of special interest to hunters and conservation groups—than about other species. If waterfowl species are known to be present along a given route, then a mistnet model could infer that suitable habitat must have been available to them. The model could then infer that the same habitat must have been

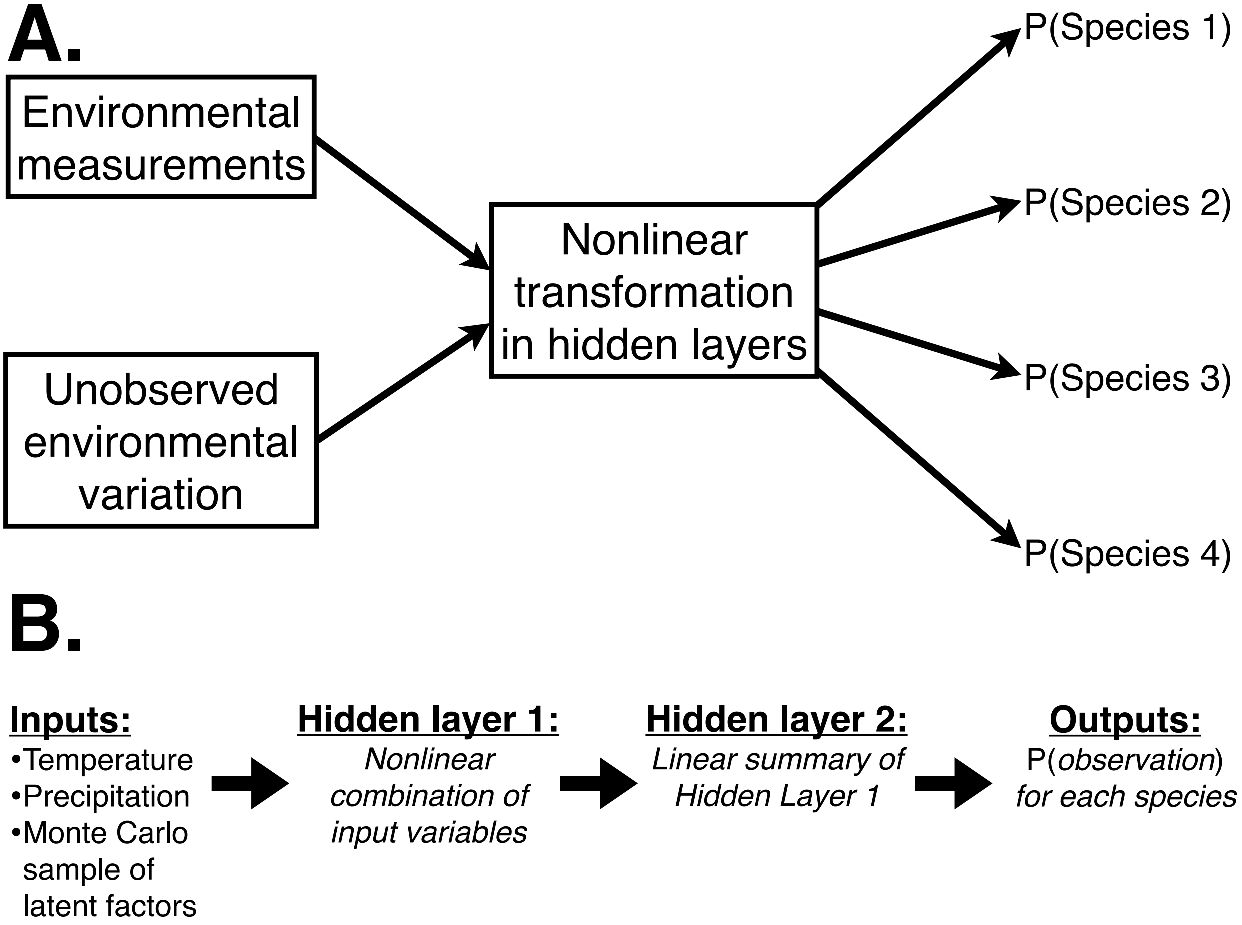

A.
Environmental measurements
Unobserved environmental variation
Nonlinear transformation in hidden layers
P(Species 1)
P(Species 2)
P(Species 3)
P(Species 4)
B.
Inputs:
•Temperature
•Precipitation
•Monte Carlo sample of latent factors
Hidden layer 1:
Nonlinear combination of input variables
Hidden layer 2:
Linear summary of Hidden Layer 1
Outputs:
P(observation) for each species

Figure 1.3: **A** A generalized diagram for stochastic feed-forward neural networks that transform environmental variables into occurrence probabilities multiple species. The network's hidden layers perform a nonlinear transformation of the observed and unobserved ("latent") environmental variables; each species' occurrence probability then depends on the state of the final hidden layer in a generalized linear fashion. **B** The specific network used in this paper, with two hidden layers. The inputs include Worldclim variables involving temperature and precipitation, as well as random draws from each of the latent environmental factors. These inputs are multiplied by a coefficient matrix and then nonlinearly transformed in the first hidden layer. The second hidden layer uses a different coefficient matrix to linearly transform its inputs down to a smaller number of variables (like Principal Components Analysis of the previous layer's activations). A third matrix of coefficients links each species' occurrence probability to each of the variables in this linear summary (like one instance of logistic regression for each species). The coefficients are all learned using a variant of the backpropagation algorithm.

available to other species, such as grebes and rails, with similar requirements. These species' predicted occurrence probabilities should thus increase automatically wherever waterfowl have been detected. Notably, the required correlations are automatically inferred from species' co-occurrence patterns, so the accuracy of these updated predictions does not depend closely on the user's ecological intuition about species' environmental tolerances.

As with most neural networks, a mistnet model's coefficients are initialized randomly, and then an optimization procedure attempts to climb the log-likelihood surface by iteratively adjusting the coefficients toward better values (i.e. gradient-based hill-climbing). In mistnet models, these adjustments are calculated with a variant of the backpropagation algorithm suggested by Tang and Salakhutdinov (2013) (described in more detail in Appendix C). The generalized expectation maximization procedure used in this variant alternates between inferring the states of the latent variables that produced the observed assemblages (via importance sampling) and updating the model's coefficients to make better predictions (via weighted backpropagation). By iteratively improving the model's estimates of the latent environmental factors and of the parameters governing species' responses to them, this procedure will eventually bring the model—with probability one—to a local maximum likelihood estimate (Neal and Hinton 1998, Tang and Salakhutdinov 2013).

In practice, most successful neural networks are regularized to avoid overfitting, meaning that they operate on a modified likelihood surface that favors reduced model complexity (Murphy 2012). In the mistnet package, regularization is formulated as prior distributions favoring smaller-magnitude parameter values over larger ones. In Bayesian terms, this means that the model maximizes the model's posterior probability rather than the likelihood (maximum a posteriori estimation); in mathematically equivalent frequentist terms (Tibshirani 1996, Murphy 2012), mistnet maximizes a constrained or penalized likelihood.

The mistnet source code can be downloaded from
https://github.com/davharris/mistnet/releases.

1.2.3 A mistnet model for bird assemblages

Mistnet models can take a variety of forms, depending on the statistical or biological problems of interest. The model used in these analyses, shown in Figure 3b, uses two hidden layers that transform the environmental data into a form that is suitable for a linear classifier; the final layer essentially performs logistic regression in this transformed space. As discussed below, this structure is designed to improve the interpretability of the model, relative to other nonlinear SDMs.

Each hidden unit ("neuron") in the first layer is sensitive to a different axis of environmental variation (e.g. one neuron could respond positively to "cold and wet" environments, while another could respond to "hot and humid" environments). The hidden units' responses are nonlinear (Appendix B), expressing the possibility that—for example—species might be more sensitive to a one-degree change in temperature from 25-26° C than to a change of the same magnitude from 19-20° C.

The second hidden layer collapses first layer's description of the environment down to a smaller number of values (e.g. 15 in this analysis; Appendix D), using a linear transformation. Thus, the network's structure ensures that each species' response to the environment can be described using a small number of coefficients (e.g., one for each of the 15 transformed environmental variables described in the second layer, plus an intercept term). The small number of coefficients and the consistency of their ecological roles across species make mistnet models highly interpretable: the coefficients linking the second hidden layer to a given species' probability of occurrence essentially describe that species' responses to the leading principal components of environmental variation (cf Vincent et al. 2010). For comparison, the boosted regression tree SDMs used below (Elith et al. 2008) have tens of thousands of coefficients per species, with entirely new interpretations for each new species' coefficients.

Apart from limiting the number of coefficients per species, two additional factors constrained the model's capacity for overfitting. First, the coefficients in each layer were constrained using weak Gaussian priors, preventing any one variable from dominating the network. Second, a very weak $\mathrm{Beta}(1.000001, 1.000001)$ prior was used to reduce the prevalence of overconfident predictions ($|\text{odds ratio}| > 10^6$). The size of each layer and optimization details were chosen by cross-validation (see Appendix D for the settings that were evaluated, along with their cross-validated likelihoods).

1.2.4 Existing models used for comparison

I compared mistnet's predictive performance with two machine learning techniques and with a linear JSDM. Each technique is described briefly below; see Appendix D for each model's settings.

The first machine learning method I used for comparison, boosted regression trees (BRT), is among the most powerful techniques available for single-species SDM (Elith et al. 2006, 2008). I trained one BRT model for each species using the `gbm` package (Ridgeway 2013) and stacked them following the recommendations in Calabrese et al. (2014).

I also used a deterministic neural network from the `nnet` package (Venables and Ripley 2002) as a baseline to assess the importance of mistnet's latent random variables. This network shares some information among species (i.e. all species' occurrence probabilities depend on the same hidden layer), but like most other multi-species SDMs (Leathwick et al. 2005, Ferrier et al. 2007) it is not a JSDM and does not explicitly model co-occurrence (Clark et al. 2013).

Finally, I trained a linear JSDM using the BayesComm package (Golding 2013, **???**) to assess the importance of mistnet's nonlinearities compared to a linear alternative that also models co-occurrence explicitly.

1.2.5 Evaluating model predictions along test routes

I evaluated mistnet's predictions both qualitatively and quantitatively. Qualitative assessments involved looking for patterns in the model's predictions and comparing them with ornithological knowledge (e.g. the AAB habitat classifications).

Each model was evaluated quantitatively on the test routes (red points in Figure 2) to assess its predictive accuracy out-of-sample. Models were scored according to their predictive likelihoods, i.e. the probabilities they assigned to various scenarios observed in the test data. Models with high likelihoods tend to produce realistic co-occurrence patterns, and should yield more biologically relevant insights about the processes underlying those patterns. Models that overfit or underfit will have lower out-of-sample likelihoods, and drawing scientific conclusions from them could be unwise. I tested each model's ability to make several kinds of predictions, ranging from the species level to predictions about the richness and composition of entire assemblages. Models that assumed species were uncorrelated should see an exponential decay in their likelihoods as the number of species increases (since the probability of making correct predictions for a set of uncorrelated species equals the product of their individual probabilities), while BayesComm and mistnet should be able to simplify the problem for larger assemblages by using correlational information.

In addition to assessing the models' overall likelihoods, I also focused on their predictions about species richness by comparing the range of possible richness values they expected along each test route with what was actually observed. For each model, I used the Poisson-binomial distribution (Hong 2013) to find confidence intervals for species richness, as described in Calabrese et al. (2014). The Poisson-binomial distribution (not to be confused with the better-known Poisson distribution for counting rare events) represents each species' occurrence as an independent Bernoulli trial with its own probability of success; the total number of successes determines the overall richness. For the two JSDMs, I calculated the confidence intervals for the appropriate mixtures of Poisson-Binomial distributions (as estimated from 1000 independent Monte Carlo samples).

1.3 Results and Discussion

1.3.1 Mistnet's view of North American bird assemblages

I began by decomposing the variance in the mistnet's species-level predictions among routes (which varied in their climate values) and residual (within-route) variation (Appendix E). On average, the residuals accounted for 30% of the variance in mistnet's predictions, suggesting that non-climate factors play a substantial role in habitat filtering.

If the non-climate factors mistnet identified were biologically meaningful, then there should be a strong correspondence between the 12 coefficients assigned to each species by mistnet and the AAB habitat classifications. A linear discriminant analysis (LDA; Venables and Ripley 2002) demonstrated such a correspondence (Figure 4). Mistnet's coefficients cleanly distinguished several groups of species by habitat association (e.g. "Grassland" species versus "Forest" species), though the model largely failed to distinguish "Marsh" species from "Lake/Pond" species and "Scrub" species from "Open Woodland" species. These results indicate that the model has identified the broad differences among communities, but that it lacks some fine-scale resolution for distinguishing among types of wetlands and among types of partially-wooded areas. Alternatively, perhaps these finer distinctions are not as salient at the scale of a 40-km transect or require more than two dimensions to represent.

While one might be able to produce a similar-looking scatterplot using ordination methods such as nonmetric multidimensional scaling (NMDS; McCune et al. 2002), the interpretation would be very different. Species' positions in an ordination plots are chosen to preserve the multivariate geometry of the data and do not usually connect to any data-generating process or to a predictive model. In Figure 4, by contrast, each species' coordinates describe the predicted slopes of its responses to two axes of environmental variation; these slopes could be used to make specific predictions about occurrence probabilities at new sites. Likewise, deviations from these predictions could be used to falsify the underlying model, without the need for expensive permutation tests or comparison with a null model. The close connection

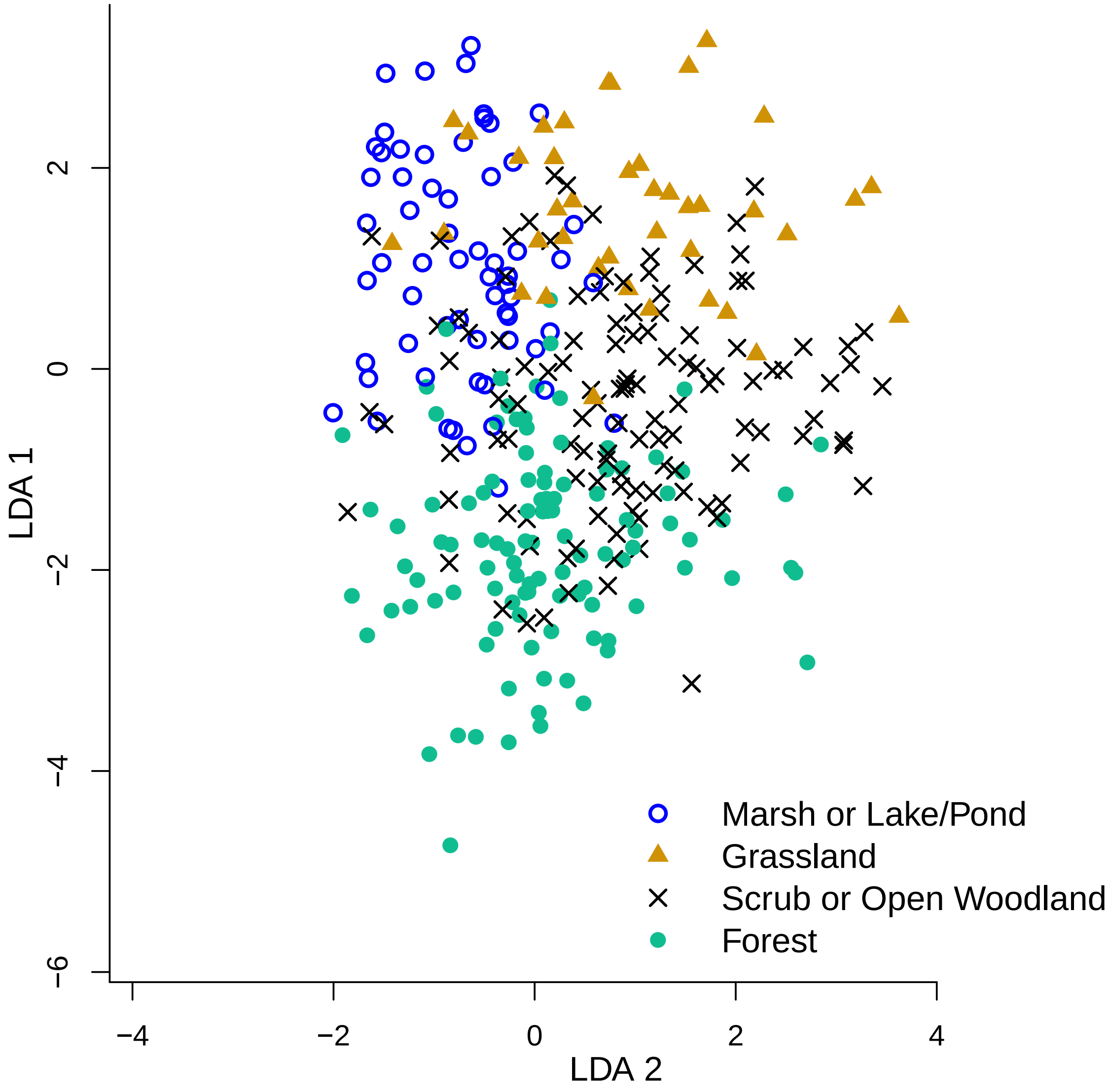


Figure 1.4: Each species' mistnet coefficients have been projected into a two-dimensional space by linear discriminant analysis (LDA), maximizing the spread between the six habitat types assigned to species by the Cornell Lab of Ornithology's All About Birds website. Mistnet cleanly separates "Grassland" species from "Forest" species, with "Scrub" and "Open Woodland" species representing intermediates along this axis of variation. "Marsh" and "Lake/Pond" species cluster together in the upper-left. Habitat classes with fewer than 15 species were omitted from this analysis.

between model and visualization demonstrated in Figure 4 may prove especially useful in contexts where prediction and understanding are both important.

The environmental gradients identified in Figure 4 are explored further in Figure 5. Figure 5A shows how the forest/grassland gradient identified by mistnet affects the model's predictions for a pair of species with opposite responses to forest cover. The model cannot tell *which* of these two species will be observed (since it was only provided with climate data), but the model has learned enough about these two species to tell that the probability of observing *both* along the same 40-km transect is much lower than would be expected if the species were uncorrelated.

Figure 5A reflects a great deal of uncertainty, which is appropriate considering that the model has no information about a crucial environmental variable (forest cover). Often, however, additional information is available that could help resolve this uncertainty, and the mistnet package includes a built-in way to do so, as indicated in Figures 5B and 5C. These panels show how the model is able to use a chance observation of a forest-associated Nashville Warbler (*Oreothlypis ruficapilla*) to indicate that a whole suite of other forest-dwelling species are likely to occur nearby, and that a variety of species that prefer open fields and wetlands should be absent. Similarly, Figure 5D shows how the presence of a Redhead duck (*Aythya americana*) can inform the model that a route likely contains suitable habitat for waterfowl, marsh-breeding blackbirds, shorebirds, and rails (along with the European Starling and Bobolink, whose true wetland associations are somewhat weaker). None of these inferences would be possible from a stack of disconnected single-species SDMs, nor would traditional ordination methods have been able to quantify the changes.

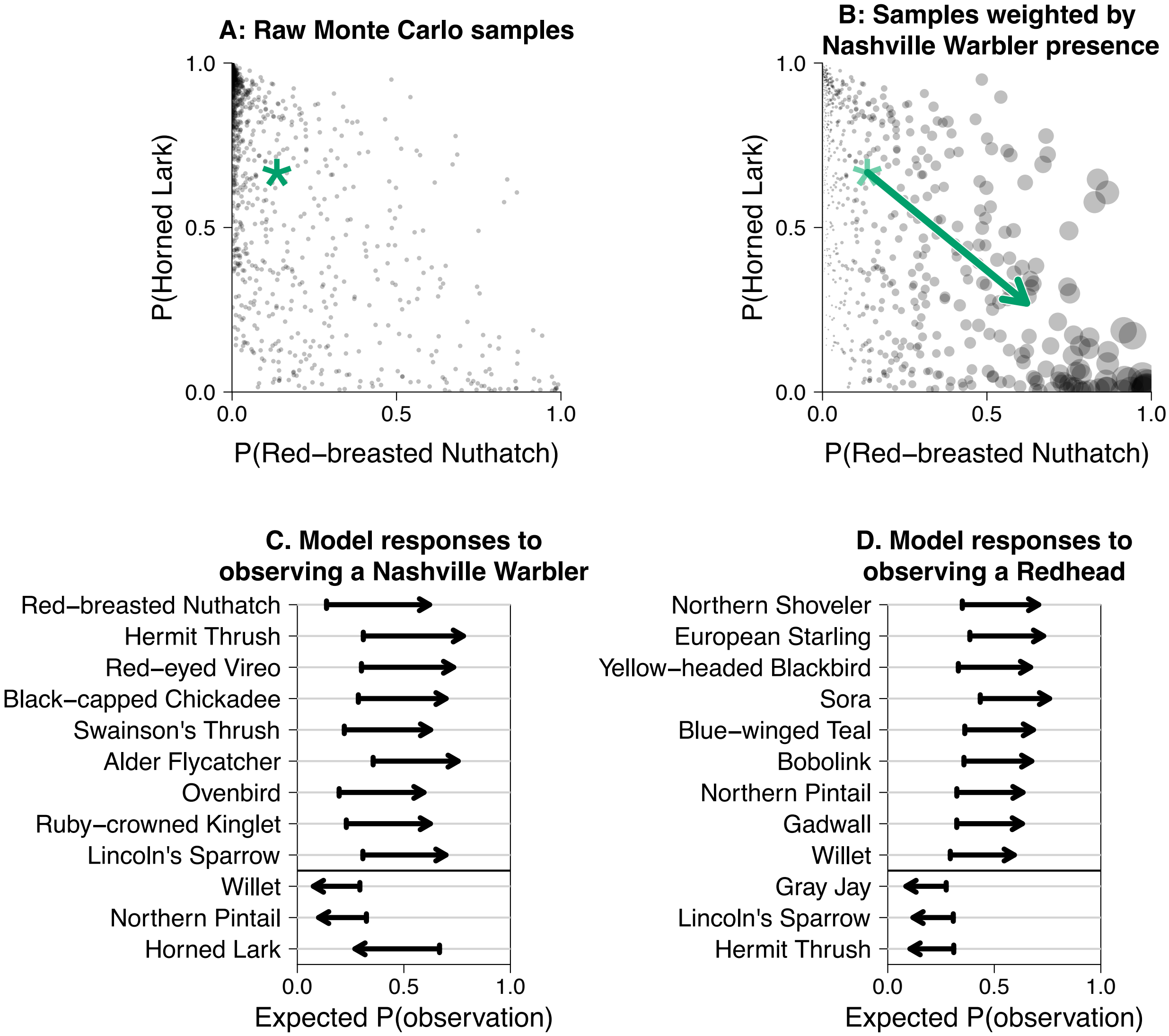


Figure 1.5: **A.** The mistnet model has learned that Red-breasted Nuthatches (*Sitta canadensis*) and Horned Larks (*Eremophila alpestris*) have opposite responses to some environmental factor whose true value is unknown. Based on these two species' biology, an ornithologist could infer that this unobserved variable is related to forest cover, with the Nuthatch favoring more forested areas and the Lark favoring more open areas. The green asterisk marks the marginal expected probability of observing the two species. **B.** The presence of a forest-dwelling Nashville Warbler (*Oreothlypis ruficapilla*) provides the model with a strong indication that the area is forested, increasing the weight assigned to Monte Carlo samples that are suitable for the Nuthatch and decreasing the weight assigned to samples that are suitable for the lark. The model's updated expectations can be found at the head of the green arrow. **C.** The Nashville Warbler's presence similarly suggests increased occurrence probabilities for a variety of other forest species (top portion of panel), and decreased probabilities for species associated with open habitat (bottom portion). **D.** If a Redhead (*Aythya americana*) had been observed instead, the model would correctly expect to see more water-associated birds and fewer forest dwellers.

1.3.2 Model comparison: species richness

Environmental heterogeneity plays an especially important role in determining species richness, which is often overdispersed relative to models' expectations (O'Hara 2005). Figure 6 shows that mistnet's predictions respect the heterogeneity one might find in nature: areas with a given climate could plausibly be either very unsuitable for most waterfowl (Anatid richness < 2 species) or much more suitable (Anatid richness > 10 species). Under the independence assumption used for stacking SDMs, however, both of these scenarios would be ruled out (Figure 6A).

Stacking leads to even larger errors when predicting richness for larger groups, such as the complete set of birds studied here. Models that stacked independent predictions consistently underestimated the range of biologically possible outcomes (Figure 6B), frequently putting million-to-one or even billion-to-one odds against species richness values that were actually observed. These models' 95% confidence intervals were so narrow that half of the observed species richness values fell outside the predicted range. The overconfidence associated with stacked models could have serious consequences in both management and research contexts if we fail to prepare for species richness values outside such unreasonably narrow bounds (e.g. expecting a reserve to protect 40-50 species even though it only supports 15). Mistnet, on the other hand, was able to explore the range of possible non-climate environments to avoid these missteps: 90% of the test routes fell within mistnet's 95% confidence intervals, and the log-likelihood ratio decisively favored it over stacked alternatives.

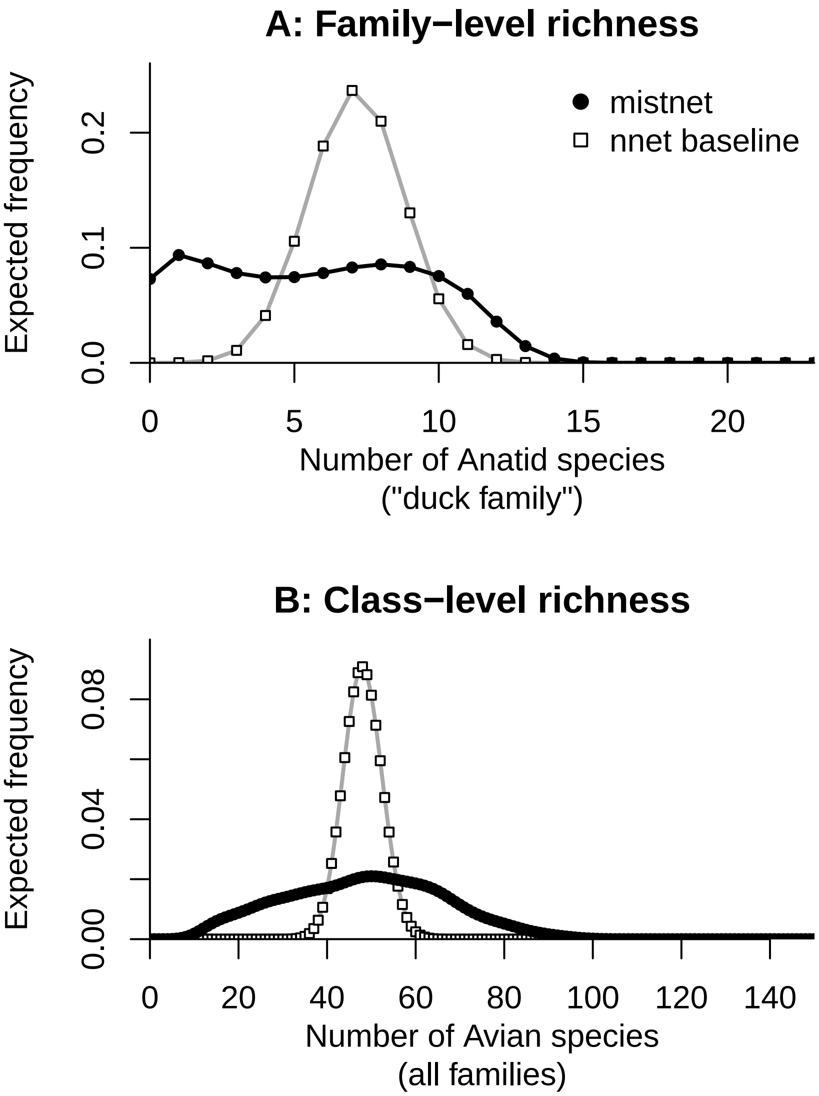


Figure 1.6: The predicted distribution of species richness values one would expect to find based on predictions from mistnet and from the deterministic neural network baseline. **A.** Anatid (waterfowl) species richness. **B.** Total species richness. BRT's predictions (not shown) are similar to the baseline network, since neither one accounts for the effects of unmeasured environmental heterogeneity. In general, both networks' mean predictions are equally distant from the observed values, but only mistnet represents its uncertainty adequately.

1.3.3 Model comparison: single species

Figure 7A compares the models' ability to make predictions for a single species across all the test routes (shown as the exponentiated expected log-likelihood). While there was substantial variation among species, the two neural network models' predictions averaged more than an order of magnitude better than BRT's. Moreover, these models' advantage over BRT was largest for low-prevalence species (linear regression of log-likelihood ratio versus log-prevalence; $p = 3 \cdot 10^{-4}$), which will often be of the greatest concern to conservationists. The most likely reason for this improvement was a reduction in overfitting: while the overall model included complex nonlinear transformations, the number of degrees of freedom associated with any given species in the final logistic regression layer was modest (15 weights plus an intercept term).

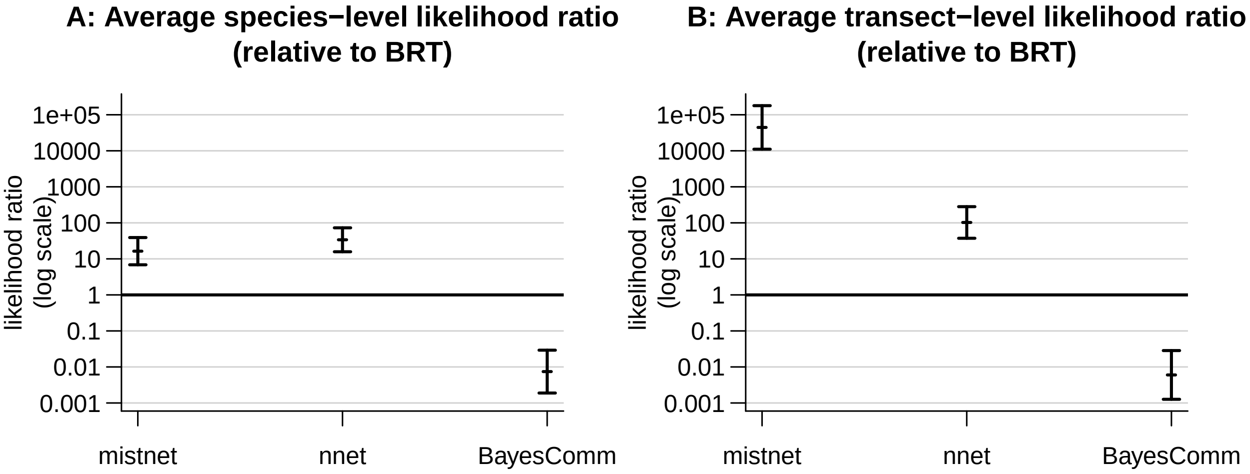


Figure 1.7: Relative predictive performance of the evaluated methods, as compared to BRT (mean +/- 95% CI, calculated from paired t-tests on the log-likelihood scale). **A.** Expected likelihood ratio for predictions about one species across 280 test-set routes. **B.** Expected likelihood ratio when predicting species composition of a test route.

BayesComm's predictions were substantially worse than any of the machine learning methods tested, which I attribute mostly to its inability to learn nonlinear responses to the environment (Elith et al. 2006). Adding quadratic terms or interaction terms (c.f. Austin 1985, Jamil and Braak 2013) would have led to severe overfitting for many rare species. Even if one added a regularizer to the software to mitigate this problem, these extra pre-specified terms may still not provide enough flexibility to compete with modern nonlinear techniques.

Applying BayesComm to a large data set also highlighted one other area where mistnet appears to outperform existing JSDMs. Despite its assumed linearity, the BayesComm model required 70,000 parameters, most of which served to to identify a distinct correlation coefficient between a single pair of species. Tracing this many parameters through hundreds of Markov chain iterations routinely caused BayesComm to run out of of memory and crash, even after the code was modified to reduce its memory footprint. Sampling long Markov chains over a dense, full-rank covariance matrix (as has apparently been done in all other JSDMs to date) thus appears to be a costly strategy with large assemblages.

1.3.4 Model comparison: community composition

While making predictions about individual species is fairly straightforward with this data set (since most species have relatively narrow breeding ranges), community ecology is more concerned with co-occurrence and related patterns involving community composition (Chase 2003). Mistnet was able to use the correlation structure of the data to reduce the number of independent bits of information needed to make an accurate prediction. As a result, mistnet's route-level likelihood averaged 430 times higher than the baseline neural network's and 45,000 times higher than BRT's (Figure 7B). BayesComm demonstrated a similar effect, but not strongly enough to overcome the low quality of its species-level predictions.

1.4 Conclusion

The large discrepancy between the performance of linear and nonlinear methods shown in Figure 7A confirms previous results: accuracy in SDM applications requires the flexibility to learn about the functional form of species' environmental responses from the data (Elith et al. 2006). Likewise, mistnet's large improvement over stacked models (Figure 6, Figure 7B) provides strong evidence that accurate assemblage-level predictions require accounting for unmeasured environmental heterogeneity—especially when reasonable confidence intervals are required. Currently, mistnet appears to be the only software package that meets both of these criteria, providing both nonlinear responses to the environment and a method for dealing with assemblage-level responses to unobserved environmental heterogeneity.

Mistnet can also identify some of the same similarities among species that a skilled biologist would expect to find. For taxa on the frontier of our knowledge, a model like mistnet could help guide the biologists to ask the best questions and organize their understanding by suggesting which species have similar habitat requirements—even when the factors controlling their occurrence are still unknown (cf. indirect gradient analysis). Unlike with stacked methods, one can read this information directly from mistnet's coefficient tables with no more difficulty than interpreting a Principal Components Analysis. Also, where most ordination techniques merely describe the multivariate geometry of an existing data matrix, mistnet's coefficients are directly tied to quantitative—and falsifiable—predictions about community structure in unobserved locations. Nonlinear JSDMs should thus be able to take on a variety of roles in ecologists' toolboxes, providing a unified framework for summarizing community structure, developing forecasts, and evaluating hypotheses about community structure.

Future research should look for ways to use other forms of ecological knowledge about species to impose some structure on models coefficients and nudge the models toward more biologically reasonable predictions (Kearney and Porter 2009, Lankau et al. 2011, Kissling et al. 2012). Such a research program could also be useful in other areas of predictive ecology (Pearse et al. 2013). JSDMs' ability to use asymmetrical or low-quality data sources to improve their predictions should also increase the value of low-effort data collection procedures such as short transects—especially since these data sources can be incorporated without the need for fitting a new model.

Finally, while it would be tempting to attribute JSDMs' correlation structure to species interactions, this approach may not be as fruitful as some authors have hoped. The correlations are all driven indirectly via shared dependencies on latent variables, instead of the direct response of one species to another implied by species interactions. Pollock et al. (2014)'s covariance decomposition allows for some progress toward inferring interactions from JSDMs, but it would be much more straightforward to use a different approach (such as Markov random fields (Azaele et al. 2010) or ensembles of classifier chains (Yu et al. 2011)) whose coefficients describe direct pairwise interactions much more explicitly. Latent variable models are more appropriate for studies like this one at large spatial scales where direct species interactions will tend to be weaker and most of the variation is driven by environmental filtering and species' range limits.

Mistnet's accuracy, interpretability, and flexibility to work with opportunistic samples indicate that nonlinear JSDMs will be important in a variety of basic and applied contexts, from forecasting, to quantifying differences among species, to developing new insights about community structure. Ecologists' models for these tasks need not be neural nets, but these analyses suggest that the most comprehensive and useful models will have many of the same features, such as latent random variables, nonlinearity, and low rank.

1.5 Acknowledgements

This work benefitted greatly from discussions with A. Sih and his lab meeting group, M. L. Baskett, R. J. Hijmans, R. McElreath, J. H. Thorne, M. W. Schwartz, B. M. Bolker, R. E. Snyder, A. C. Perry, and C. S. Tysor, as well as comments from S. C. Walker and an anonymous reviewer. It was funded by a Graduate Research Fellowship from the National Science Foundation, the UC Davis Center for Population Biology, and the California Department of Water Resources. I gratefully acknowledge the field biologists that collected the BBS data, as well as the US Geological Survey, Cornell Lab of Ornithology, and Worldclim for making their data publicly available.

1.6 Data Accessibility:

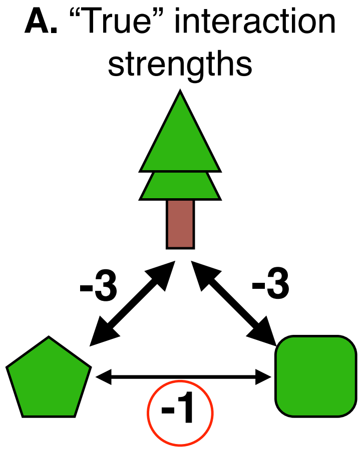
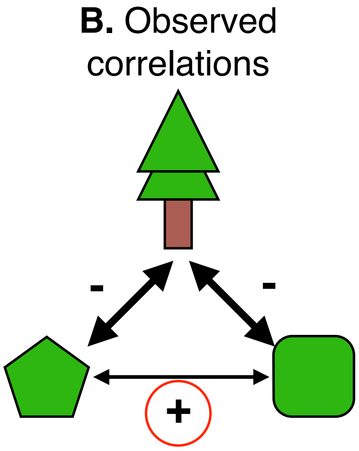
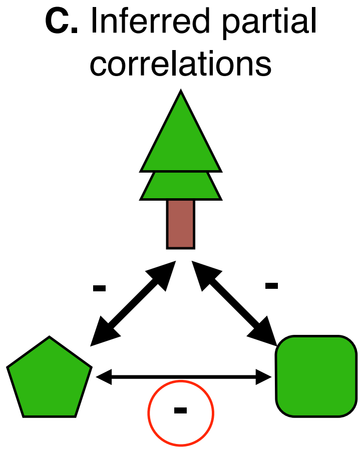
- All data sets used here are freely downloadable from their original sources.
- The mistnet source code is at https://github.com/davharris/mistnet and can be installed with the `install_github` function from the `devtools` package. The specific version used in this paper is at https://github.com/davharris/mistnet/releases/tag/v0.2.0

2 Estimating species interactions from observational data with Markov networks

David J. Harris

2.1 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, Gotelli and Entsminger 2003, Connor et al. 2013), existing methods for detecting competition's effects on community structure are unreliable (Gotelli and Ulrich 2009). In particular, species' effects on one another can become lost in the complex web of direct and indirect interactions in real assemblages. For example, the competitive interaction between the two shrub species in Figure 1A can become obscured by their shared tendency to occur in unshaded areas (Figure 1B). While ecologists have long known that indirect effects can overwhelm direct ones at the landscape level (Dodson 1970, Levine 1976), the vast majority of our methods for drawing inferenes from observational data do not control for these effects (e.g. Diamond 1975, Strong et al. 1984, Gotelli and Ulrich 2009, Veech 2013, Pollock et al. 2014). To the extent that indirect interactions like those in Figure 1 are generally important (Dodson 1970), existing methods will thus not generally provide much evidence regarding species' direct effects on one another. The goal of this paper is to resolve this long-standing problem.

A. “True” interaction strengths
-3
-3
-1
B. Observed correlations
-
-
+
C. Inferred partial correlations
-
-
-

Figure 2.1: **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 the tree species' presence with a conditional method such as a partial covariance or a Markov network leads to correct identification of the negative shrub-shrub interaction (circled).

While competition doesn't reliably reduce co-occurrence rates at the whole-landscape level (as most of our methods assume), it nevertheless does leave a signal in the data (Figure 1C). Specifically, after partitioning the data set into shaded sites and unshaded sites, there will be co-occurrence deficits in each subset that might not be apparent at the landscape level. More generally, we can obtain much better estimates of the association between two species from their conditional relationships (i.e. by controlling for other species in the network) than we could get from their overall co-occurrence rates. This kind of precision is difficult to obtain from null models, which begin with the assumption that all the pairwise interactions are zero and thus don't need to be controlled for. Nevertheless, null models have dominated this field for more than three decades (Strong et al. 1984, Gotelli and Ulrich 2009).

Following recent work by Azaele et al. (2010) and Fort (2013), this paper shows that Markov networks (undirected graphical models also known as Markov random fields; Murphy 2012) can provide a framework for understanding the landscape-level consequences of pairwise species interactions, and for detecting them from observed presence-absence matrices. Markov networks have been used in many scientific fields in similar contexts for decades, from physics (where nearby particles interact magnetically; Cipra 1987) to spatial statistics (where adjacent grid cells have correlated values; Harris 1974, Gelfand et al. 2005). While community ecologists explored some related approaches in the 1980's (Whittam and Siegel-Causey 1981), they used severe approximations that led to unintelligible results (e.g. "probabilities" greater than one; Gilpin and Diamond 1982).

Below, I introduce Markov networks and show how they can be used to simulate landscape-level data or to make exact predictions about the direct and indirect consequences of possible interaction matrices. Then, using simulated data sets where the “true” interactions are known, I compare this approach with several existing methods. Finally, I discuss opportunities for extending the approach presented here to other problems in community ecology, e.g. quantifying the overall effect of species interactions on occurrence rates (Roughgarden 1983) and disentangling the effects of biotic versus abiotic interactions on species composition (Kissling et al. 2012, Pollock et al. 2014).

2.2 Methods

2.2.1 Markov networks.

Markov networks provide a framework for translating back and forth between the conditional relationships among species (Figure 1C) and the kinds of species assemblages that these relationships produce. Here, I show how a set of conditional relationships can be used to determine how groups of species can co-occur. Methods for estimating conditional relationships from data are discussed in the next section.

A Markov network defines the relative probability of observing a given vector of species-level presences (1s) and absences (0s), $\vec{y}$, as

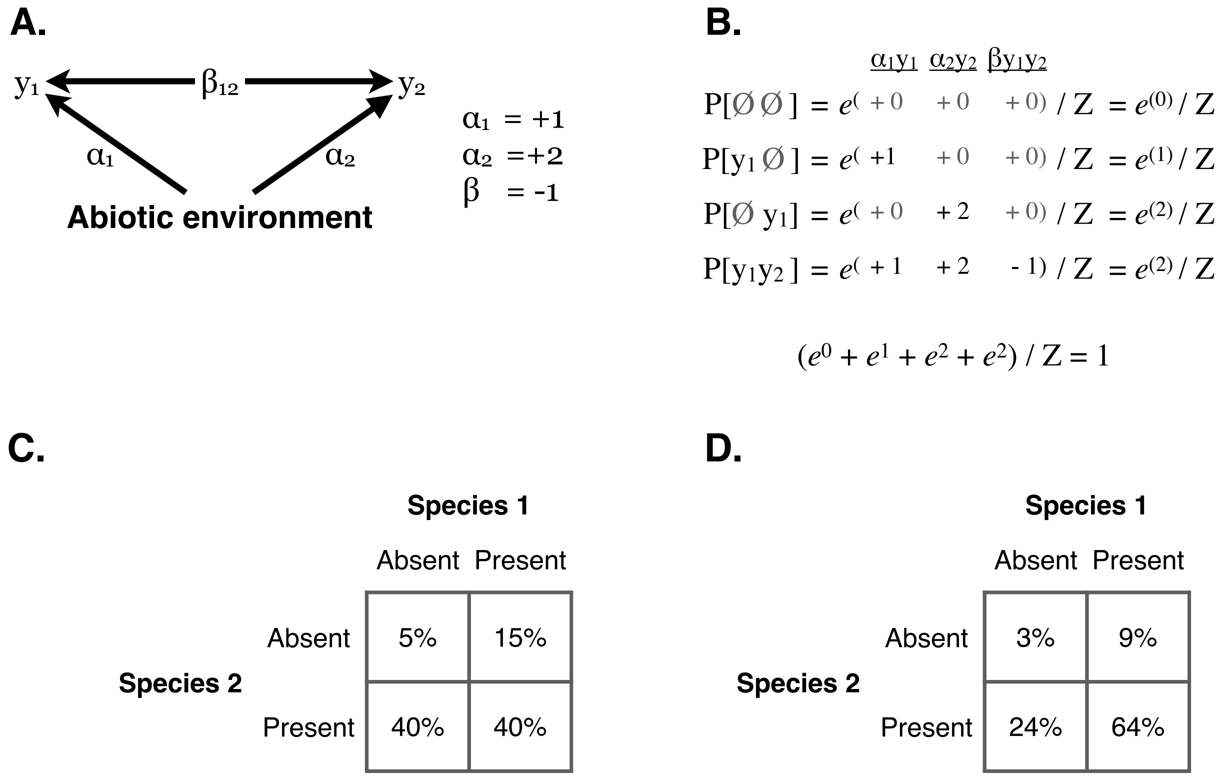
$$p(\vec{y}; \alpha, \beta) \propto exp(\textstyle\sum_i \alpha_i y_i + \sum_{i \neq j} \beta_{ij} y_i y_j).$$

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 conditional relationship between two species, i.e. the probability that they will be found together, after controlling for the other species in the network (Figure 2A, Figure 2B). For example, β_{ij} might have a value of +2 for two mutualists, indicating that the odds of observing one species are e^2 times higher in sites where its partner is present than in comparable sites where its partner is absent. 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—all else equal—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).

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. Determining the marginal relationships between species from their conditional interactions entails summing over the different possible assemblages (Figure 2B). This becomes intractable when the number of possible assemblages is large, though several methods beyond the scope of this paper can be employed to keep the calculations feasible (Salakhutdinov 2008, Lee and Hastie 2012). Alternatively, as noted below, some common linear and generalized linear methods can also be used as computationally efficient approximations to the full network (Lee and Hastie 2012, Loh and Wainwright 2013).

2.2.2 Estimating α and β coefficients from presence-absence data.

In the previous section, the values of α and β were known and the goal was to make predictions about possible species assemblages. In practice, however, ecologists will often need to estimate the parameters from an observed co-occurrence matrix (i.e. from a matrix



$$\begin{array}{llll} & \underline{\alpha_1 y_1} & \underline{\alpha_2 y_2} & \underline{\beta y_1 y_2} \\ P[\emptyset\,\emptyset] = e^{(} & +0 & +0 & +0)\ /\ Z\ = e^{(0)}/\ Z \\ P[y_1\,\emptyset] = e^{(} & +1 & +0 & +0)\ /\ Z\ = e^{(1)}/\ Z \\ P[\emptyset\ y_1] = e^{(} & +0 & +2 & +0)\ /\ Z\ = e^{(2)}/\ Z \\ P[y_1 y_2] = e^{(} & +1 & +2 & -1)\ /\ Z\ = e^{(2)}/\ Z \end{array}$$

$$(e^0 + e^1 + e^2 + e^2)\ /\ Z = 1$$

		Species 1	
		Absent	Present
Species 2	Absent	5%	15%
	Present	40%	40%

		Species 1	
		Absent	Present
Species 2	Absent	3%	9%
	Present	24%	64%

Figure 2.2: **Figure 2. A.** A small Markov network with two species. The abiotic environment favors the occurrence of both species ($\alpha > 0$), particularly species 2 ($\alpha_2 > \alpha_1$). The negative β coefficient linking these two species implies that they co-occur less than expected under independence. **B.** Relative probabilities of all four possible presence-absence combinations for Species 1 and Species 2. The exponent includes α_1 whenever Species 1 is present ($y_1 = 1$), but not when it is absent ($y_1 = 0$). Similarly, the exponent includes α_2 only when species 2 is present ($y_2 = 1$), and β only when both are present ($y_1 y_2 = 1$). The normalizing constant Z, ensures that the four relative probabilities sum to 1. In this case, Z is about 18.5. **C.** We can find the expected frequencies of all possible co-occurrence patterns between the two species of interest. **D.** If β_{12} 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 co-occurrence deficit would be eliminated.

of ones and zeros indicating which species are present at which sites). When the number of species is reasonably small, one can compute exact maximum likelihood estimates for all of the α and β coefficients given a presence-absence matrix 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 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 matches the sufficient statistics of the model distribution to 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. The package was named after Santa Rosalia, the patron saint of biodiversity, whose supposedly miraculous healing powers played an important rhetorical role in the null model debates of the 1970's and 1980's (Lewin 1983).

2.2.3 Simulated landscapes.

In order to compare different methods, I simulated two sets of landscapes using known parameters. The first set 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 also simulated a second set of landscapes using a stochastic community model based on generalized Lotka-Volterra dynamics, as described in Appendix 2. In these simulations, each species pair was randomly assigned to either compete for a portion of the available carrying capacity (negative interaction) or to act as mutualists (positive interaction). Here, mutualisms operate by mitigating the effects of intraspecific competition on each partner's death rate. For these analyses, I simulated landscapes with up to 20 species and 25, 200, or 1600 sites (50 replicates per landscape size; see Appendix 2).

2.2.4 Recovering species interactions from simulated data.

I compared seven techniques for determining the sign and strength of the associations between pairs of species from simulated data (Appendix 3). First, I used the rosalia package (Harris 2015a) to fit Markov newtork models, as described above. For the analyses with 20 species, I added a very weak logistic prior distribution on the α and β terms with scale 2 to ensure that the model estimates were always finite. The bias introduced by this prior should be small: the 95% credible interval on β only requires that one species' effect on the odds of observing a different species to be less than a factor of 1500 (which is not much of a constraint). The logistic distribution was chosen because it is convex and has a similar shape to the Laplace distribution used in LASSO regularization (especially in the tails), but unlike the Laplace distribution it is differentiable everywhere and does not force any estimates to be exactly zero. To confirm that this procedure produced stable estimates, I compared its estimates on 50 bootstrap replicates (Appendix 4).

I also evaluated six alternative methods: five from the existing literature, plus a novel combination of two of these methods. The first alternative interaction metric was the sample correlation between species' presence-absence vectors, which summarizes their marginal association. Next, I used partial correlations, which summarize species' conditional relationships (Albrecht and Gotelli 2001, Faisal et al. 2010). In the context of non-Gaussian data, the partial correlation can be thought of as a computationally efficient approximation to the full Markov network model (Loh and Wainwright 2013). This sort of model is very common for estimating relationships among genes and gene products (Friedman et al. 2008). Because partial correlations are undefined for landscapes with perfectly-correlated species pairs, I used a regularized estimate based on James-Stein shrinkage, as implemented in the corpcor package's `pcor.shrink` function with the default settings (Schäfer et al. 2014).

The third alternative, generalized linear models (GLMs), can also be thought of as a computationally 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. To avoid the identifiability problems associated with directed cyclic graphs (Schmidt and Murphy 2012), I then symmetrized the relationships within species pairs via averaging.

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.

2.2.5 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 and noted whether these intervals included zero.

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. Because the different methods mostly describe species interactions on different scales (e.g. correlations versus Z scores versus regression coefficients), I used linear regression through the origin to rescale the different estimates produced by each method so that they had a consistent interpretation. After rescaling each method's estimates, I calculated squared errors between the scaled interaction estimates and "true" interaction values across all the simulated data sets. These squared errors determined 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).

2.3 Results

2.3.1 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).

2.3.2 Twenty species.

Despite some variability across contexts (Figure 3A), the four methods that controlled for indirect effects clearly performed the best: the Markov network explained the largest portion of the variance in the "true" interaction coefficients (35% overall), followed by the generalized linear models (30%), partial correlations from the raw presence-absence data (28%), and partial correlations from BayesComm, the joint species distribution model (26%). The benefit of choosing the full Markov network over the other three methods was largest on the smaller landscapes, which are also the ones that are most representative of typical analyses in this field (Gotelli and Ulrich 2009).

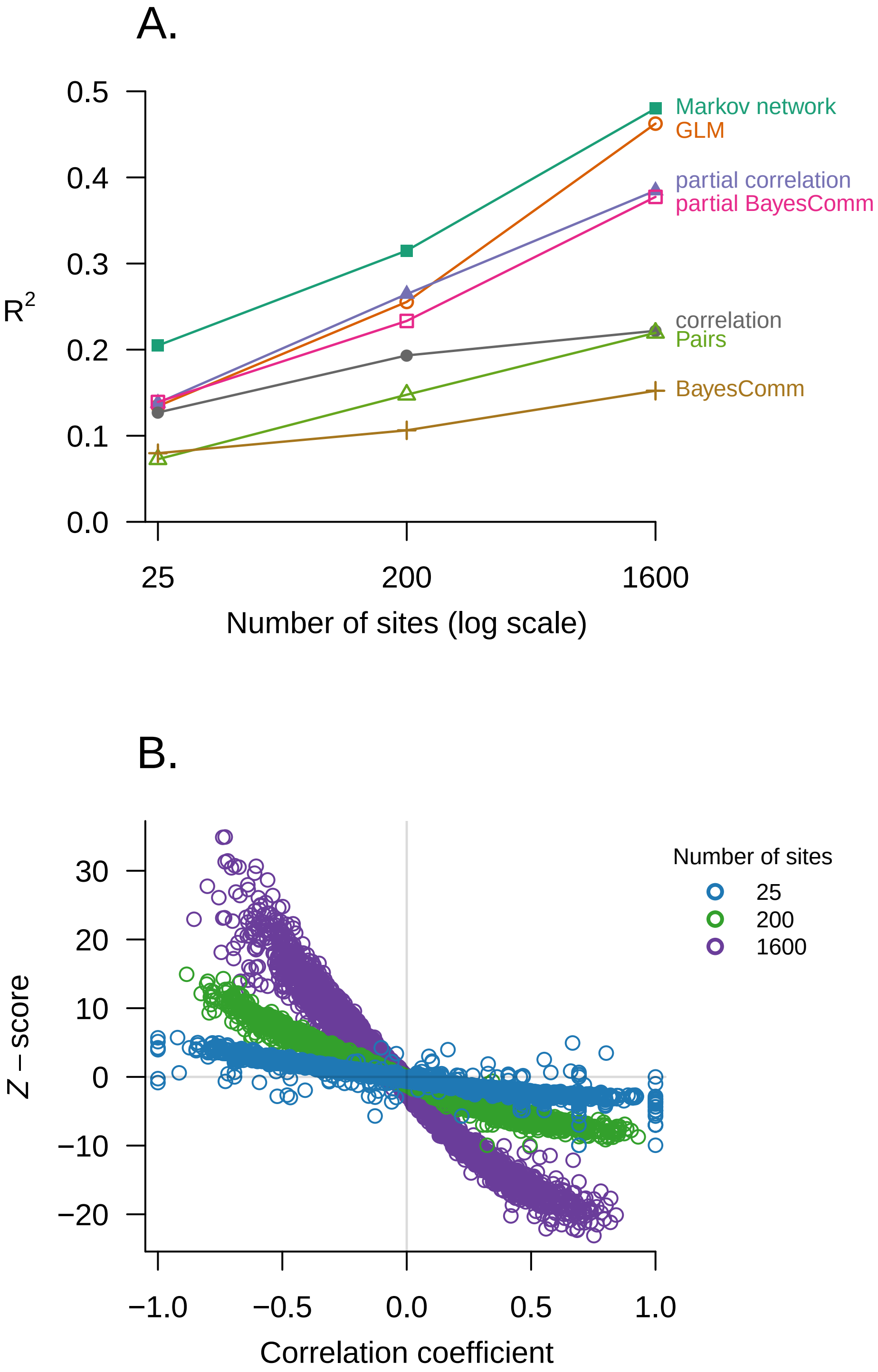


Figure 2.3: **Figure 3. A.** Proportion of variance in interaction coefficients explained by each method versus number of sampled locations. **B.** The Z-scores produced by the null model ("Pairs") for each pair of species can be predicted using the correlation between the presence-absence vectors of those same species and from the number of sites on the landscape.

The three methods that did not attempt to control for indirect interactions all explained less than 20% of the variance. Of these, the sample correlation matrix based on the raw data performed the best (19%), followed by the null model (15%) and BayesComm's correlation matrix (11%). Although these last three methods had different R^2 values, there was a close mapping among their estimates (especially after controlling for the size of the simulated landscapes; Figure 3B). This suggests that the effect sizes from the null model (and, to a lesser extent, the correlation matrices from joint species distribution models) only contain noisy versions of the same information that could be obtained more easily and interpretably by calculating correlation coefficients between species' presence-absence vectors.

Bootstrap resampling indicated that the above ranking of the different methods was robust (Appendix 3). In particular, the 95% confidence interval of the bootstrap distribution indicated that the Markov network's overall R^2 value was between 14 and 18 percent higher than the second-most effective method (generalized linear models) and between 2.12 and 2.38 times higher than could be achieved by the null model (Pairs). Bootstrap resampling of a 200-site landscape also confirmed that the rosalia package's estimates of species' conditional relationships were robust to sampling variation for reasonably-sized landscapes (Appendix 4).

2.4 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). For cases where fitting a

Markov network is computationally infeasible, these results also indicate that partial covariances and generalized linear models (the two methods that estimated conditional relationships rather than marginal ones) can both provide useful approximations. 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). Similarly, the approximation involved in using separate generalized linear models for each species can occasionally lead to catastrophic overfitting with small-to-moderate sample sizes (Lee and Hastie 2012). For these reasons, it will usually be best to fit a Markov network rather than one of the alternative methods when one's computational resources allow it.

It's important to note that none of these methods can identify the exact nature of the pairwise interactions (e.g. which species in a positively-associated pair is facilitating the other; Schmidt and Murphy 2012), particularly when real pairs of species can reciprocally influence one another in multiple ways simultaneously (Bruno et al. 2003); with compositional data, there is only enough information to provide a single number describing each species pair. To estimate asymmetric interactions, such as commensalism or predation, ecologists would need other kinds of data, as from time series, behavioral observations, manipulative experiments, or natural history. These other sources of information could also be used to augment the likelihood function with an informative prior distribution, which could lead to better results on some real data sets than was shown in Figure 3A.

Despite their limitations, Markov networks have enormous potential to improve ecological understanding. In particular, they are less vulnerable than some of the most commonly-used methods to mistakenly identifying positive species interactions between competing species, and can make precise statements about the conditions where indirect interactions will overwhelm direct ones. They also provide a simple answer to the question of how competition should affect a species' overall prevalence, which was a major flashpoint 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 ($\frac{e^{\alpha}}{1+e^{\alpha}}$; Lee and Hastie 2012). Competition's effect on prevalence in a Markov network can then be calculated by subtracting this value from the observed prevalence (cf Figure 2D). This kind of insight would have been difficult to obtain without a generative model that makes predictions about the consequences of species interactions; null models (which presume *a priori* that interactions do not exist) have no way to make such predictions.

Markov networks—particularly the Ising model for binary networks—have been studied for nearly a century (Cipra 1987), and the models' properties, capabilities, and limits are well-understood in a huge range of applications. Using the same framework for species interactions 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. Numerous other extensions are possible: for example, the states of the interaction network can be modeled as a function of the local abiotic environment (Lee and Hastie 2012), which would provide a rigorous and straightforward approach to the difficult and important task of incorporating whole networks of biotic interactions into species distribution models (Kissling et al. 2012, Pollock et al. 2014), leading to a better understanding of the interplay between biotic and abiotic effects on community structure. There are even methods (Whittam and

Siegel-Causey 1981, Tjelmeland and Besag 1998) that would allow one species to affect the sign or strength of the relationship between two other species, tipping 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 for studying species interactions. Null and neutral models can be 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 particular null model must be interpreted with care (Roughgarden 1983). 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), especially when the test statistic is effectively just a correlation coefficient (Figure 3B). Simultaneous estimation of multiple ecological parameters 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 3A, scientists in this field could often double their explanatory power by switching from null models to Markov networks (or increase it substantially with linear or generalized linear approximations). Regardless of the methods ecologists ultimately choose, controlling for indirect effects could clearly improve our understanding of species' direct effects on one another and on community structure.

2.5 Acknowledgements:

This research was funded by a Graduate Research Fellowship from the US National Science Foundation and benefited greatly from discussions with A. Sih, M. L. Baskett, R. McElreath,

R. J. Hijmans, A. C. Perry, and C. S. Tysor. Additionally, A. K. Barner, E. Baldridge, E. P. White, D. Li, D. L. Miller, N. Golding, N. J. Gotelli, C. F. Dormann, and two anonymous reviewers provided very helpful feedback on the text.

References

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

Austin, M. P. 1985. Continuum concept, ordination methods, and niche theory. Annual Review of Ecology and Systematics:39–61.

Austin, M. P., and K. P. Van Niel. 2011. Improving species distribution models for climate change studies: Variable selection and scale. Journal of Biogeography 38:1–8.

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.

Bahn, V., and B. J. McGill. 2007. Can niche-based distribution models outperform spatial interpolation? Global Ecology and Biogeography 16:733–742.

Bengio, Y. 2013. Deep Learning of Representations: Looking Forward. Pages 1–37 *in* A.-H. Dediu, C. Martín-Vide, R. Mitkov, and B. Truthe, editors. Statistical Language and Speech Processing. Springer Berlin Heidelberg.

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

Calabrese, J. M., G. Certain, C. Kraan, and C. F. Dormann. 2014. Stacking species distribution models and adjusting bias by linking them to macroecological models. Global Ecology and Biogeography 23:99–112.

Chase, J. M. 2003. Community assembly: When should history matter? Oecologia 136:489–498.

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

Clark, J. S., A. E. Gelfand, C. W. Woodall, and K. Zhu. 2013. MORE THAN THE SUM OF THE PARTS: FOREST CLIMATE RESPONSE FROM JOINT SPECIES

DISTRIBUTION MODELS. Ecological Applications.

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.

Dormann, C. F., O. Purschke, J. R. G. Márquez, S. Lautenbach, and B. Schröder. 2008. Components of uncertainty in species distribution analysis: A case study of the great grey shrike. Ecology 89:3371–3386.

Elith, J., and J. R. Leathwick. 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology, Evolution, and Systematics 40:677–697.

Elith, J., C. H. Graham*, R. P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, J. Li, L. G. Lohmann, B. A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. McC. M. Overton, A. Townsend Peterson, S. J. Phillips, K. Richardson, R. Scachetti-Pereira, R. E. Schapire, J. Soberón, S. Williams, M. S. Wisz, and N. E. Zimmermann. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129–151.

Elith, J., J. R. Leathwick, and T. Hastie. 2008. A working guide to boosted regression trees. Journal of Animal Ecology 77:802–813.

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.

Ferrier, S., G. Manion, J. Elith, and K. Richardson. 2007. Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. Diversity and Distributions 13:252–264.

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

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. 2013. PhD thesis: Mapping and understanding the distributions of potential vector mosquitoes in the UK: New methods and applications. figshare.

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

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.

Guisan, A., and C. Rahbek. 2011. SESAM – a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. Journal of Biogeography 38:1433–1444.

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.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis. 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25:1965–1978.

Hong, Y. 2013. Poibin: The Poisson Binomial Distribution.

Hutchinson, R. A., L.-P. Liu, and T. G. Dietterich. 2011. Incorporating boosted regression trees into ecological latent variable models. Pages 1343–1348 *in* Twenty-Fifth AAAI Conference on Artificial Intelligence.

Jackson, M. M., M. G. Turner, S. M. Pearson, and A. R. Ives. 2012. Seeing the forest and the trees: Multilevel models reveal both species and community patterns. Ecosphere 3:art79.

Jamil, T., and C. J. ter Braak. 2013. Generalized linear mixed models can detect unimodal species-environment relationships. PeerJ 1:e95.

Kearney, M., and W. Porter. 2009. Mechanistic niche modelling: Combining physiological and spatial data to predict species' ranges. Ecology letters 12:334–350.

Kissling, W. D., C. F. Dormann, J. Groeneveld, T. Hickler, I. Kühn, G. J. McInerny, J. M. Montoya, C. Römermann, K. Schiffers, F. M. Schurr, A. Singer, J.-C. Svenning, N. E. Zimmermann, and R. B. O'Hara. 2012. Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. Journal of Biogeography 39:2163–2178.

Lankau, R., P. S. Jørgensen, D. J. Harris, and A. Sih. 2011. Incorporating evolutionary principles into environmental management and policy. Evolutionary Applications 4:315–325.

Latimer, A. M., S. Banerjee, H. Sang Jr, E. S. Mosher, and J. A. Silander Jr. 2009. Hierarchical models facilitate spatial analysis of large data sets: A case study on invasive plant species in the northeastern United States. Ecology Letters 12:144–154.

Leathwick, J. R., D. Rowe, J. Richardson, J. Elith, and T. Hastie. 2005. Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. Freshwater Biology 50:2034–2052.

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.

McCune, B., J. B. Grace, and D. L. Urban. 2002. Analysis of ecological communities. MjM software design Gleneden Beach, OR.

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

Neal, R. M. 1992. Connectionist learning of belief networks. Artificial Intelligence 56:71–113.

Neal, R. M., and G. E. Hinton. 1998. A view of the EM algorithm that justifies incremental, sparse, and other variants. Pages 355–368 *in* Learning in graphical models. Springer.

Ovaskainen, O., J. Hottola, and J. Siitonen. 2010. Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. Ecology 91:2514–2521.

O'Hara, R. B. 2005. Species richness estimators: How many species can dance on the head of a pin? Journal of Animal Ecology 74:375–386.

Pearse, I. S., D. J. Harris, R. Karban, and A. Sih. 2013. Predicting novel herbivore–plant interactions. Oikos 122:1554–1564.

Pellissier, L., A. Espíndola, J.-N. Pradervand, A. Dubuis, J. Pottier, S. Ferrier, and A. Guisan. 2013. A probabilistic approach to niche-based community models for spatial forecasts of assemblage properties and their uncertainties. Journal of Biogeography 40:1939–1946.

Pollock, L. J., R. Tingley, W. K. Morris, N. Golding, R. B. O'Hara, K. M. Parris, P. A. Vesk, and M. A. McCarthy. 2014. Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). Methods in Ecology and Evolution:n/a–n/a.

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

Ridgeway, G. 2013. Gbm: Generalized Boosted Regression Models.

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

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

Sauer, J. R., J. E. Hines, J. Fallon, K. Pardieck, D. Ziolkowski Jr, and W. Link. 2011. The North American breeding bird survey, results and analysis 1966-2011. Version 2011.0.

Schäfer, J., R. Opgen-Rhein, V. Zuber, M. Ahdesmäki, A. P. D. Silva, and K. Strimmer. 2014. Corpcor: Efficient Estimation of Covariance and (Partial) Correlation.

Schmidt, M., and K. Murphy. 2012. Modeling Discrete Interventional Data using Directed Cyclic Graphical Models. arXiv preprint arXiv:1205.2617.

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

Tang, Y., and R. Salakhutdinov. 2013. Learning Stochastic Feedforward Neural Networks. Pages 530–538 *in* Advances in Neural Information Processing Systems 26.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological):267–288.

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

Veech, J. A. 2013. A probabilistic model for analysing species co-occurrence. Global Ecology and Biogeography 22:252–260.

Venables, W. N., and B. D. Ripley. 2002. Modern Applied Statistics with S. Fourth. Springer, New York.

Vincent, P., H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. The Journal of Machine Learning Research 9999:3371–3408.

Walker, S. C., and D. A. Jackson. 2011. Random-effects ordination: Describing and predicting multivariate correlations and co-occurrences. Ecological Monographs 81:635–663.

Wang, Y., U. Naumann, S. T. Wright, and D. I. Warton. 2012. Mvabund–an R package for model-based analysis of multivariate abundance data. Methods in Ecology and Evolution 3:471–474.

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

Wisz, M. S., J. Pottier, W. D. Kissling, L. Pellissier, J. Lenoir, C. F. Damgaard, C. F. Dormann, M. C. Forchhammer, J.-A. Grytnes, A. Guisan, R. K. Heikkinen, T. T. Høye, I. Kühn, M. Luoto, L. Maiorano, M.-C. Nilsson, S. Normand, E. Öckinger, N. M. Schmidt, M. Termansen, A. Timmermann, D. A. Wardle, P. Aastrup, and J.-C. Svenning. 2013. The role of biotic interactions in shaping distributions and realised assemblages of species:

Implications for species distribution modelling. Biological Reviews 88:15–30.

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

Yu, J., W.-K. Wong, T. Dietterich, J. Jones, M. Betts, S. Frey, S. Shirley, J. Miller, and M. White. 2011. Multi-label Classification for Multi-Species Distribution Modeling. Proceedings of the 28th International Conference on Machine Learning.