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:

_____

Andrew Sih, Chair

_____

Marissa L. Baskett

_____

Robert Hijmans

_____

Richard McElreath

Committee in charge

2015

i

Contents

0.1 Abstract

asdf

0.2   Acknowledgements

asdf

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 pre-specified 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.
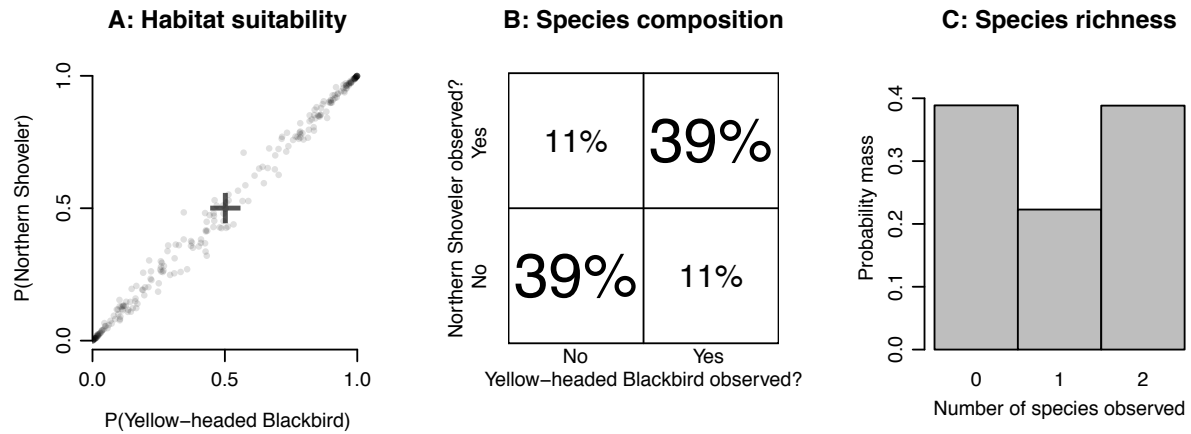
Figure 1.1: Unobserved environmental heterogeneity can induce correlations between species; ignoring this heterogeneity can produce misleading results. **A**: Based on climate predictors, a pair of single-species models might predict 50% occurrence probabilities for each of two wetland species (black 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.
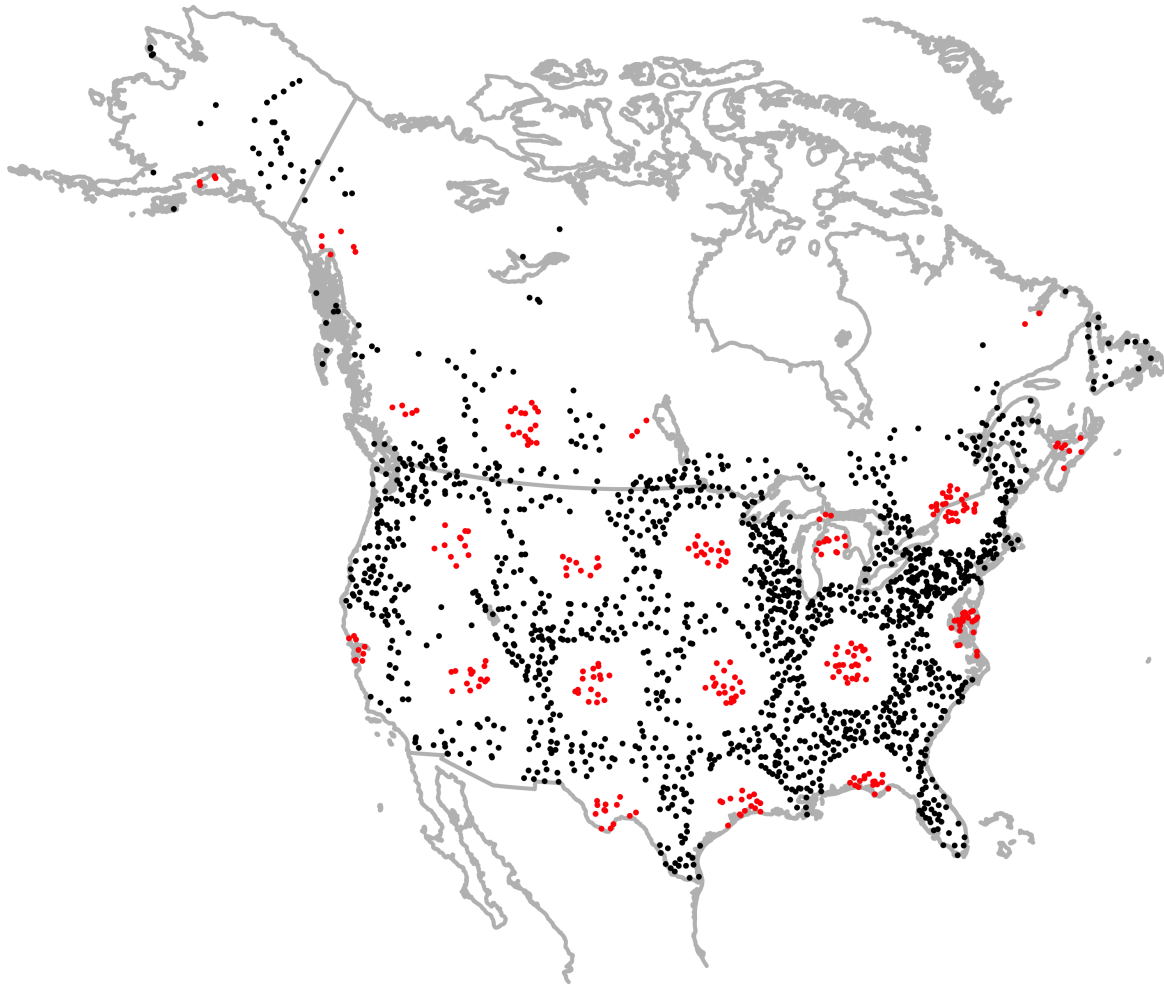
Figure 1.2: Map of the BBS routes used in this analysis. Black points are training routes; red ones are test routes. The training and test routes are separated by a 150-km buffer in order to minimize spatial autocorrelation across the two partitions.

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.

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*
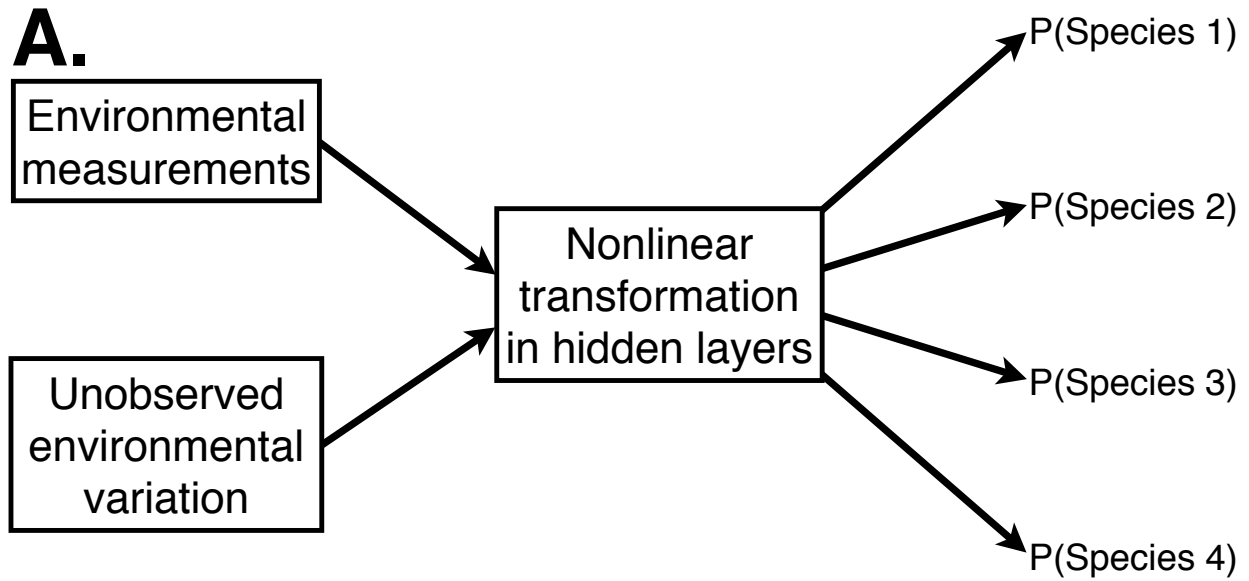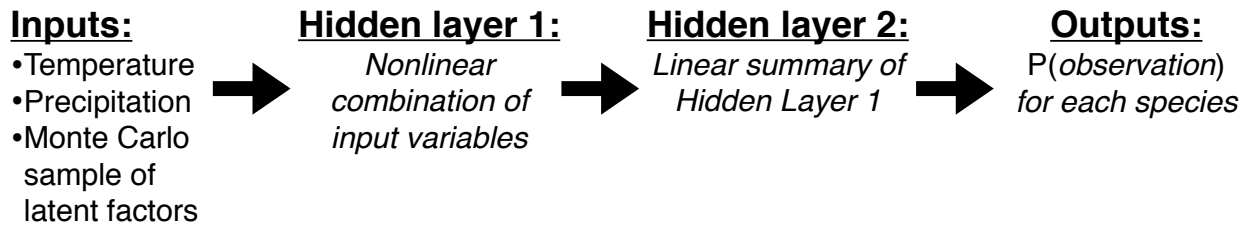
Figure 1.3: **A** A generalized diagram for stochastic feed-forward neural networks that transform environmental variables into occurrence probabilities multiple species. The network's hidden layers perform a nonlinear transformation of the observed and unobserved ("latent") environmental variables; 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

11

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 $\text{Beta}(1.000001, 1.000001)$ 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, Golding and Harris 2015) 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,

13

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
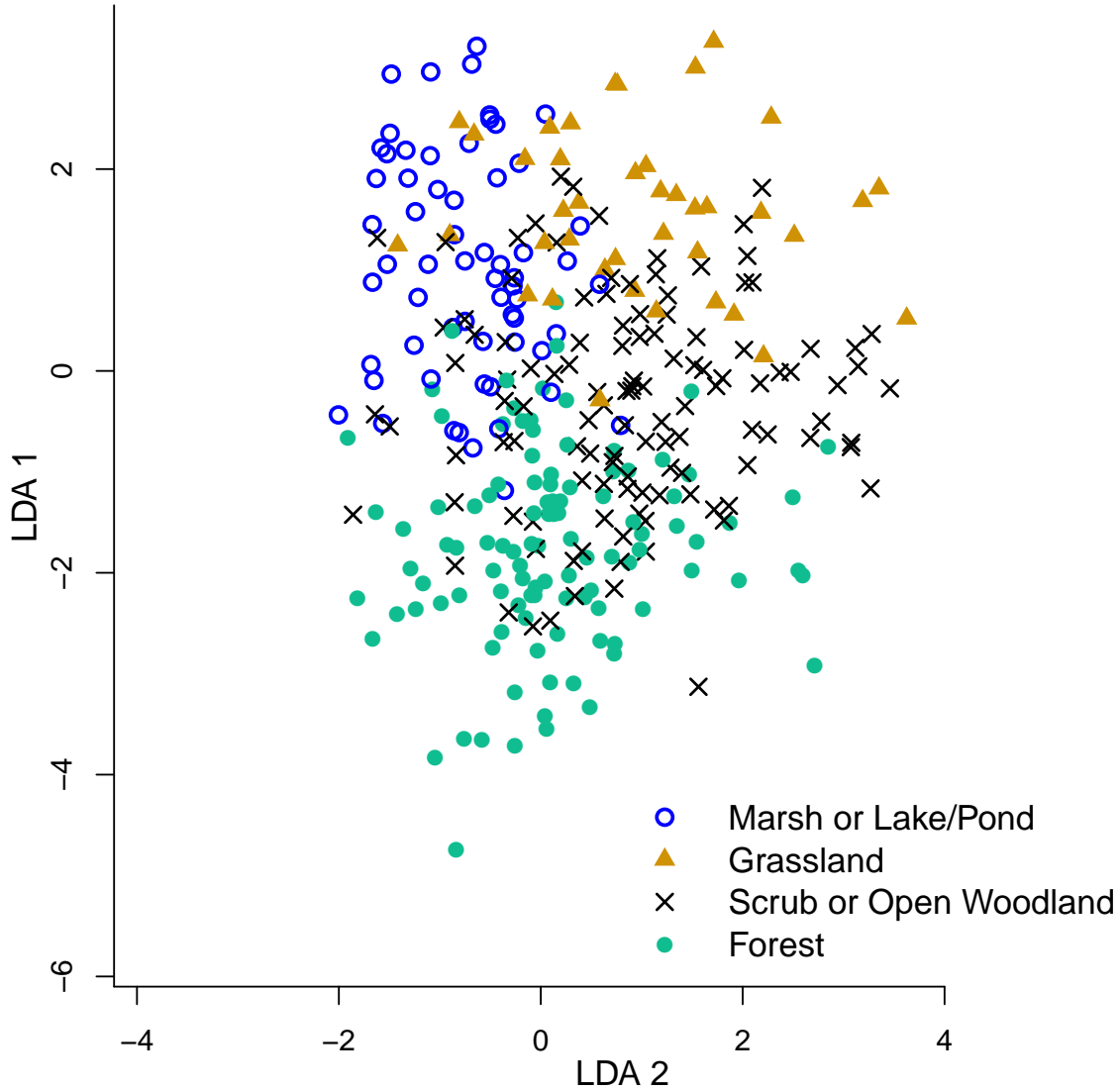
Figure 1.4: Each species' mistnet coefficients have been projected into a two-dimensional space by linear discriminant analysis (LDA), maximizing the spread between the six habitat types assigned to species by the Cornell Lab of Ornithology's All About Birds website. Mistnet cleanly separates "Grassland" 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.

A: Raw Monte Carlo samples

P(Horned Lark) vs P(Red−breasted Nuthatch)

B: Samples weighted by Nashville Warbler presence

P(Horned Lark) vs P(Red−breasted Nuthatch)

C. Model responses to observing a Nashville Warbler

- Red−breasted Nuthatch
- Hermit Thrush
- Red−eyed Vireo
- Black−capped Chickadee
- Swainson's Thrush
- Alder Flycatcher
- Ovenbird
- Ruby−crowned Kinglet
- Lincoln's Sparrow
- Willet
- Northern Pintail
- Horned Lark

Expected P(observation)

D. Model responses to observing a Redhead

- Northern Shoveler
- European Starling
- Yellow−headed Blackbird
- Sora
- Blue−winged Teal
- Bobolink
- Northern Pintail
- Gadwall
- Willet
- Gray Jay
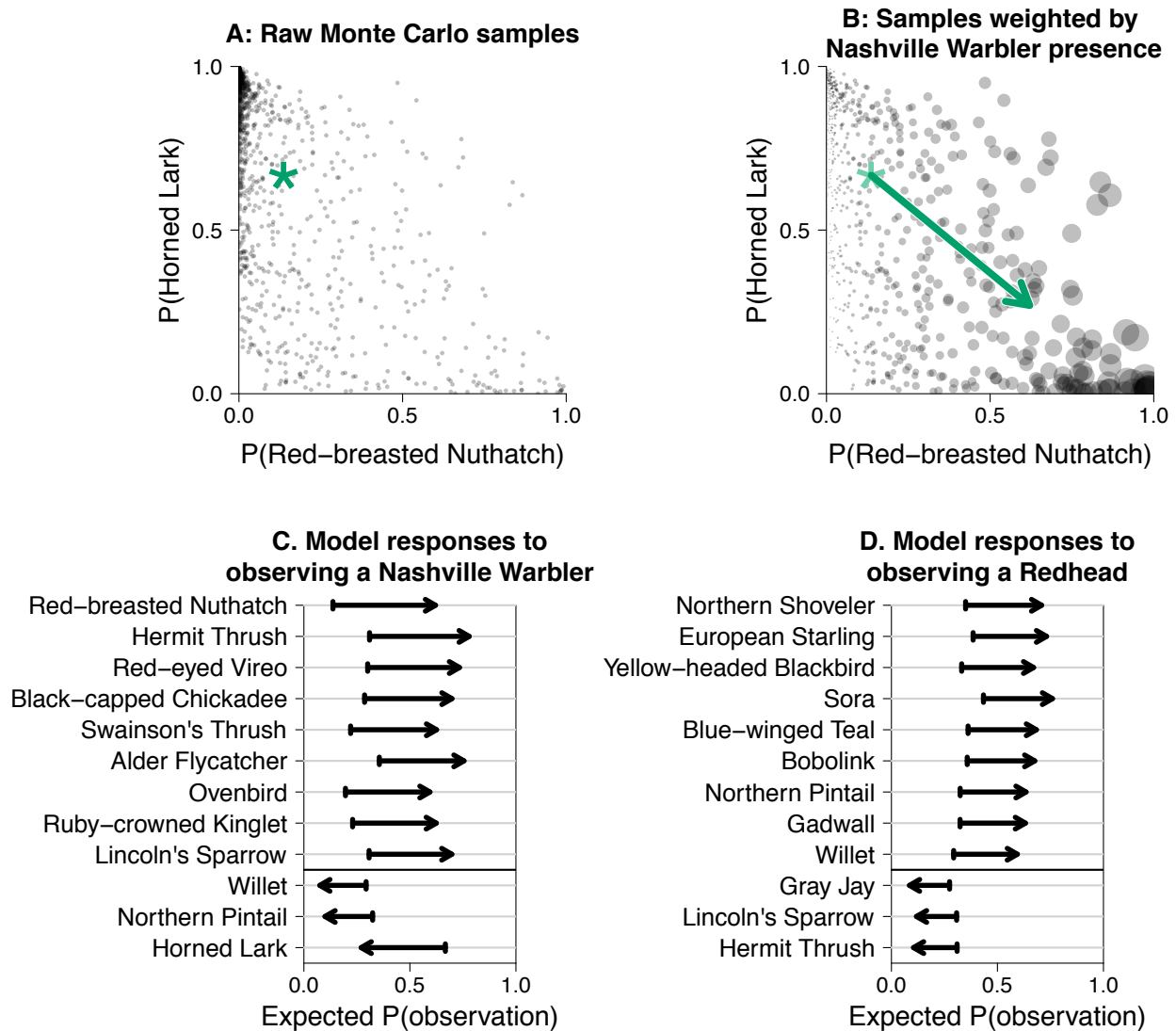- Lincoln's Sparrow
- Hermit Thrush

Expected P(observation)

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.
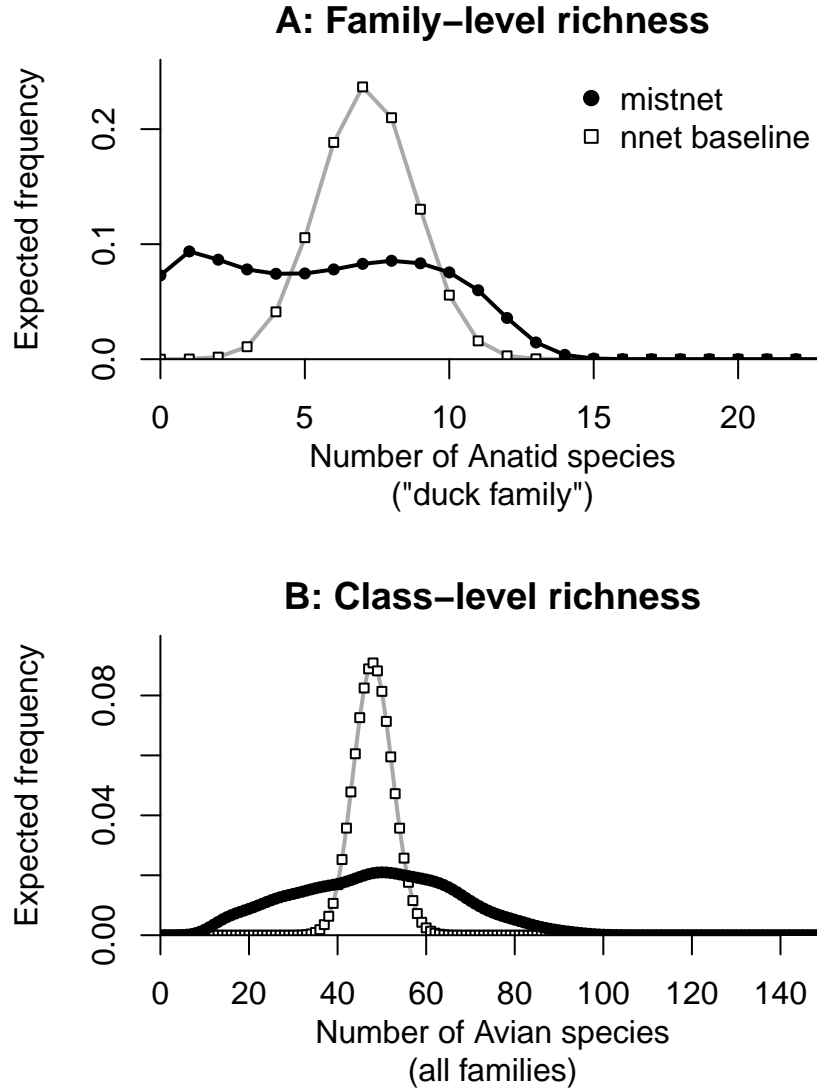
Figure 1.6: The predicted distribution of species richness values one would expect to find based on predictions from mistnet and from the deterministic neural network baseline. **A.** Anatid (waterfowl) species richness. **B.** Total species richness. BRT's predictions (not shown) are similar to the 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).
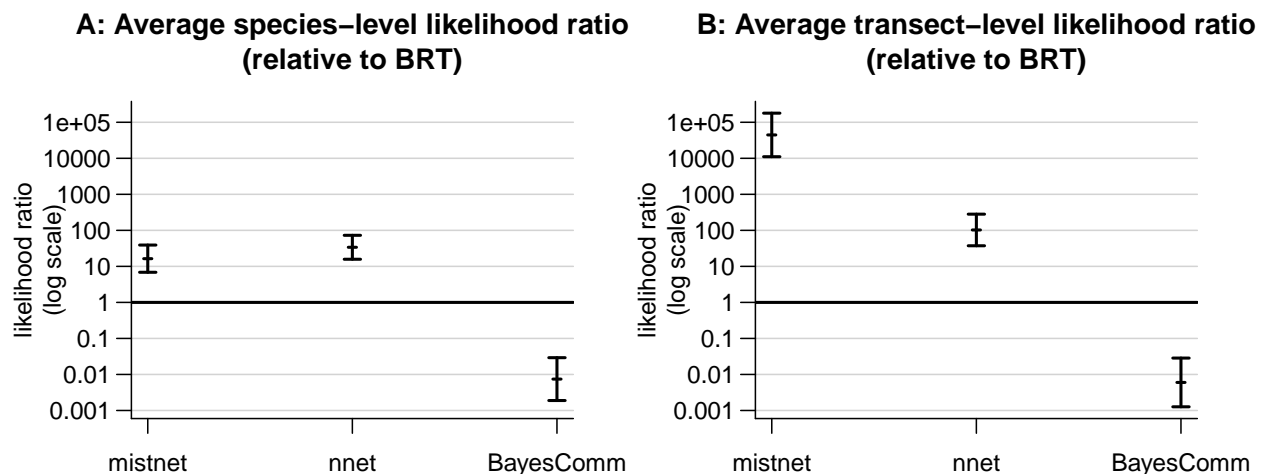

Figure 1.7: Relative predictive performance of the evaluated methods, as compared to BRT (mean +/- 95% CI, calculated from paired t-tests on the log-likelihood scale). **A.** Expected likelihood ratio for predictions about one species across 280 test-set routes. **B.** Expected likelihood ratio when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

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.
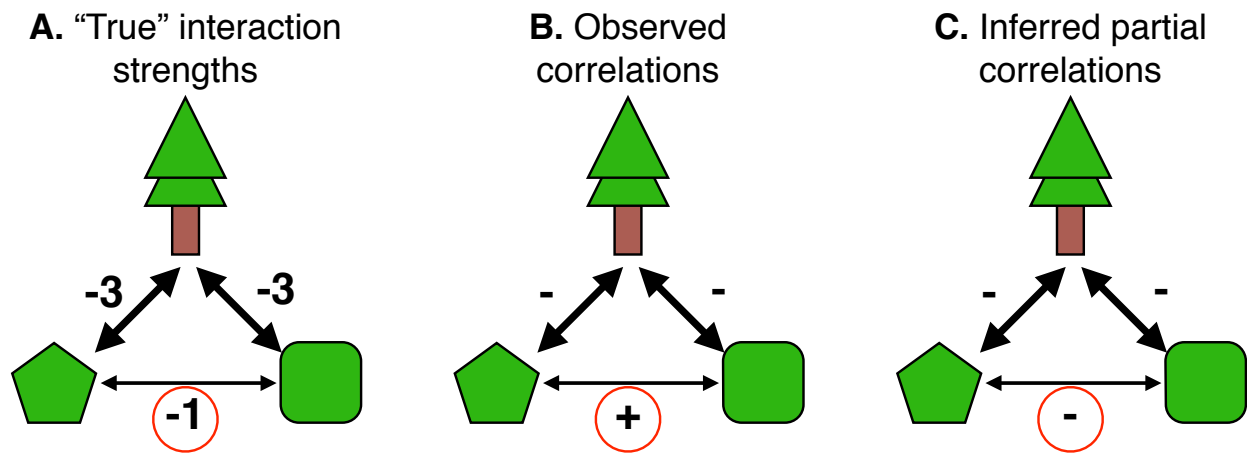
Figure 2.1: **Figure 1. A.** A small network of three competing species. The tree (top) tends not to co-occur with either of the two shrub species, as indicated by the strongly negative coefficient linking them. The two shrub species also compete with one another, but more weakly (circled coefficient). **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

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

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

29

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

A.

y_1 ⟷ β_{12} ⟶ y_2

α_1 α_2

Abiotic environment

$\alpha_1 = +1$
$\alpha_2 = +2$
$\beta = -1$

B.
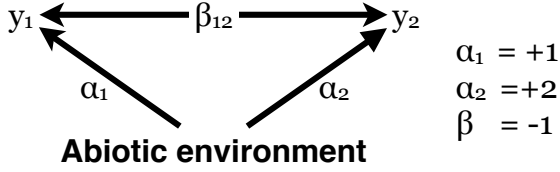
$$\alpha_1 y_1 \quad \alpha_2 y_2 \quad \beta y_1 y_2$$

$$P[\varnothing\,\varnothing] = e^{(\ +0\ \ +0\ \ +0)}/Z = e^{(0)}/Z$$

$$P[y_1\,\varnothing] = e^{(\ +1\ \ +0\ \ +0)}/Z = e^{(1)}/Z$$

$$P[\varnothing\,y_1] = e^{(\ +0\ \ +2\ \ +0)}/Z = e^{(2)}/Z$$

$$P[y_1 y_2] = e^{(\ +1\ \ +2\ \ -1)}/Z = e^{(2)}/Z$$

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

C.

Species 1

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

D.

Species 1

		Absent	Present
	Absent	3%	9%
Species 2	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).
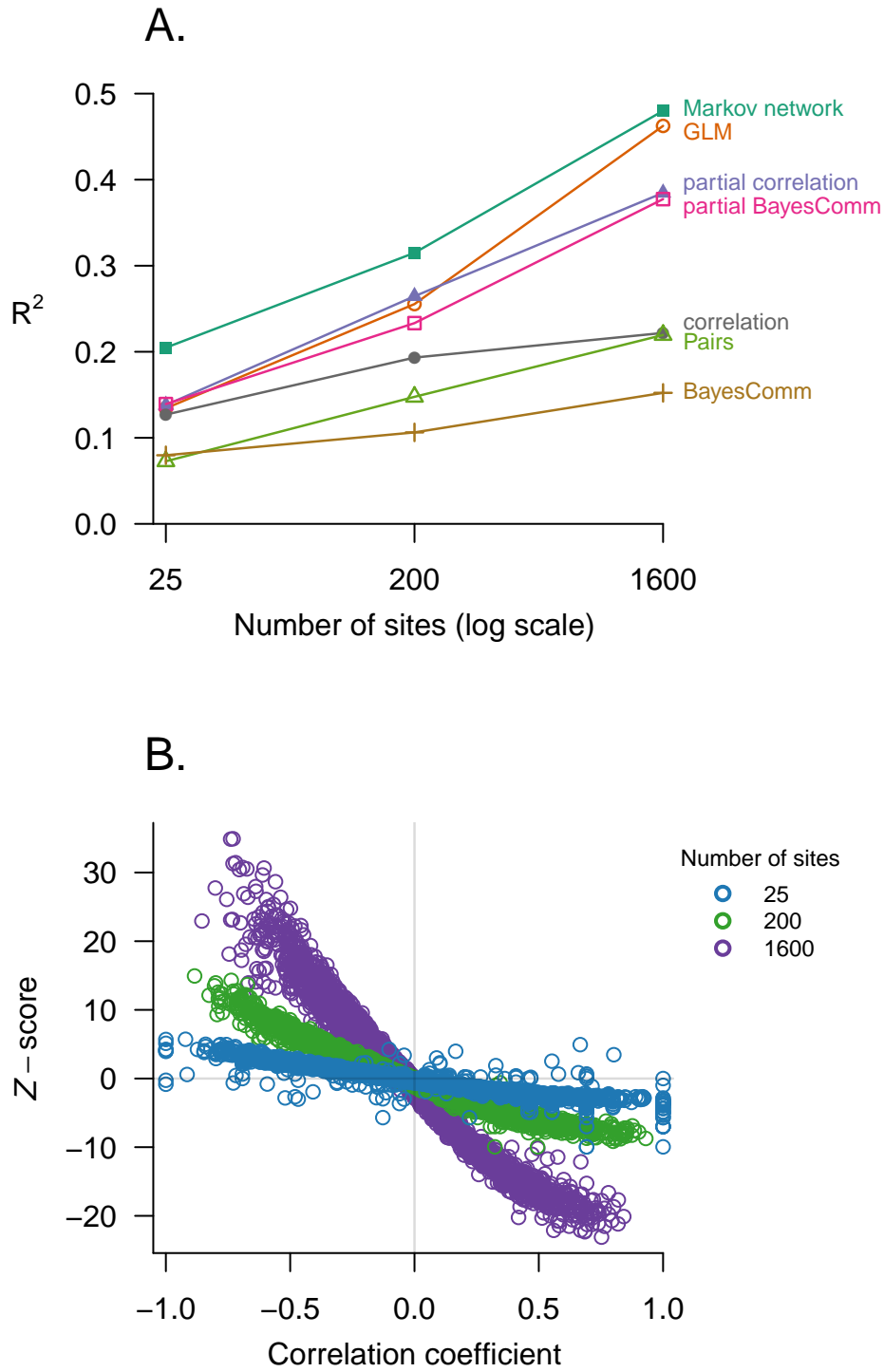
Figure 2.3: **Figure 3. A.** Proportion of variance in interaction coefficients explained by each method versus number of sampled locations. **B.** The Z-scores produced by the null model ("Pairs") for each pair of species can be predicted using the correlation between the presence-absence vectors of 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 ecolegal 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:

3 Estimating species interactions in large, abiotically structured communities with Markov networks and stochastic approximation

David J. Harris

3.1 Introduction

To the extent that species interactions are important for community assembly, ecologists generally expect them to leave a signature on species' co-occurrence patterns. Using that signature to infer the underlying species interactions from observational data has been much harder, however. Disagreements about how to draw these inferences led to an acrimonious, decade-long argument among community ecologists (Lewin 1983), where essentially all of the proposed statistical approaches were criticized for having high error rates, a poor match with the underlying ecological questions, computational infeasibility, or all three (Connor and Simberloff 1979, Gilpin and Diamond 1982, Strong et al. 1984, Hastings 1987). As documented below, ecologists have, for the most part, not solved these problems in the ensuing decades. Resolving these issues would solve a longstanding issue in community ecology and make a significant contribution to other fields, such as species distribution modeling (Kissling et al. 2012).

During the time that ecologists have struggled to identify the interactions among dozens of species, bioinformaticians have largely solved the analogous problem of estimating interactions among thousands of genes from their co-expression levels (e.g. Friedman et al. (2008)). The discrepancy is due to the fact that ecologists have chosen to focus, almost

exclusively, on overall co-occurrence rates (Connor and Simberloff 1979, Gilpin and Diamond 1982, Gotelli and Ulrich 2009, Veech 2013) or on closely-related values such as correlation coefficients (Pollock et al. 2014; see Faisal et al. 2010 and Harris 2015 for two recent exceptions). As ecologists know, however, the correlation between two species will often reflect other factors beyond their direct pairwise interactions (e.g. abiotic influences and indirect biotic effects; Figure 1). When these factors are not accounted for, this reliance on overall co-occurrence rates can reliably lead to incorrect inferences (Harris 2015c). While some methods have been proposed for incorporating a small number of specific factors (such as geographic or environmental dissimilarity) into ecologists' null models at a time (Lessard et al. 2011), we still lack a good way to scale these approaches up for cases with many biotic and abiotic factors acting simultaneously.

A number of procedures have been developed for controlling for the influence of extraneous factors and focusing on the direct relationship between a single pair of variables, i.e. for estimating the conditional relationship between them. The most familiar of these is the partial correlation. Rather than describing the overall relationship between two variables across all conditions, partial correlations (like regression coefficients) describe the portion of the relationship that remains after the other variables in the data have been accounted for. Harris (2015c) found that the partial correlation can do extract accurate information about pairwise species interactions from binary co-occurrence data in some circumstances, but the fact that this approach assumes multivariate Gaussian data makes it less appealing.

Recently, several papers have suggested that ecologists could use Markov networks (undirected graphical models also known as Markov random fields) to estimate conditional
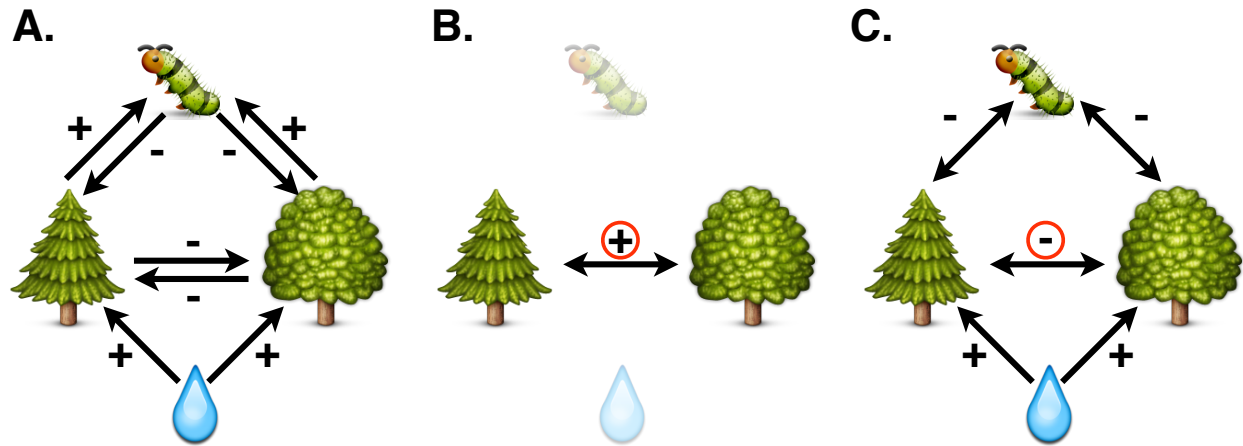
Figure 3.1: A hypothetical network of three species, two of which depend on an abiotic factor (annual rainfall). **A.** The "true" network involves the caterpillar exploiting the two tree species as well as competition between the two trees. The two trees also benefit from rainfall. **B.** Although the two trees directly reduce one another's occurrence probabilities via competition, they tend to occur in the same areas (with sufficient water and without too many herbivores). An analysis that ignores these other factors could mistakenly conclude that the two trees were mutualists because of their high co-occurrence rate (circled "+" sign). **C.** The methods presented here are designed to estimate the direct interactions between each pair of species, after accounting for other species and abiotic factors that could affect their co-occurrence rates (note that the circled interaction between the trees has the correct sign). Note that, with observational data, the bi-directional effects of species interactions (pairs of arrows between species) need to be collapsed to a single number describing the conditional relationship between species (one double-headed arrow).

relationships from binary data by maximum entropy or maximum likelihood (Azaele et al. (2010); Harris (2015c)). A Markov network defines a probability distribution over possible binary species assemblages, and its coefficients (including one coefficient describing the conditional relationship between each pair of species in the network). This approach is optimal in the sense that it produces a model that matches the observed properties of the data (i.e. the occurrence and co-occurrence rates), and that it does so with the fewest possible parameters or constraints (i.e. the Markov network has the most information entropy of any possible model satisfying its constraints; Azaele et al. 2010). As shown in Harris (2015c), Markov networks do a better job of estimating species' influences on one another better than a number of existing methods, particularly when sample sizes are low.

Unfortunately, Markov networks have an intractable likelihood function whose computational difficulty more than doubles each time a new species is added to the model. This exponential growth in the likelihood function's complexity means that the methods that worked for Azaele et al. (2010) and Harris (2015c) with networks of 20 species would be completely infeasible for networks with 50 species, requiring over a billion times more computational effort ($2^{50}/2^{20} \approx 10^9$). Extending the method to account for abiotic variation among sites on the landscape would require repeating these expensive computations independently for every site, increasing the computational burden even further. If ecologists want to apply these methods to larger problems, a different model-fitting algorithm would be needed.

In this paper, I present a different way of optimizing the likelihood, called "stochastic approximation" (Robbins and Monro 1951, Salakhutdinov and Hinton 2012), which replaces

the intractable computations with tractable Monte Carlo estimates of the same quantities. Despite the sampling error introduced by this substitution, stochastic approximation provides strong guarantees for eventual convergence to the maximum likelihood estimate (Younes 1999, Salakhutdinov and Hinton 2012). This change in approach makes it feasible to estimate the interactions among hundreds of species from observational data while also accounting for possible responses to abiotic factors (such as those used in species distribution models), giving ecologists some power to distinguish between species pairs that co-occur due to shared environmental tolerances from those that co-occur due to direct interactions like mutualism.

3.2   Methods

3.2.1   Markov networks

As discussed in Azaele et al. (2010) and Harris (2015c), Markov networks such as the Ising model (Cipra 1987) can be used to describe community structure as follows. Each species is represented by a binary random variable, describing whether it is present (1) or absent (0). A species' conditional probability of presence under a given set of biotic and abiotic conditions depends on its coefficients. The coefficients linking species to one another describe their conditional associations: all else equal, a species will occur less often in the presence of its competitors and other species with which it is negatively associated, and more often in the presence of its mutualists:

$$p(y_i|\vec{y}_{j\neq i}) = \text{logistic}\left(\alpha_i + \sum_{j\neq i}\beta_{ij}y_j\right),$$

where y_i is 1 when species i is present and 0 when it is absent, α_i is an intercept term, β_{ij} represents the conditional relationship between species i and species j, and logistic$(x) = e^x/(1 + e^x)$. These conditional probabilities can be combined to calculate the probability of observing any given combination of presences and absences, up to a multiplicative constant (Lee and Hastie 2012, Murphy 2012):

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

Unfortunately, the exponentially large number of possible assemblages makes the normalizing constant for this joint probability distribution intractable when the number of species is larger than 20 or so. However, it is generally straightforward to simulate examples of assemblages that are consistent with graphical models such as these, using Monte Carlo techniques like Gibbs sampling (Salakhutdinov and Hinton 2012). Gibbs sampling is especially convenient for these models because it only requires the conditional probabilities, which are straightforward to compute (Harris 2015; Appendix).

In each of the scenarios below, the "observed" landscapes were simulated using Gibbs sampling from a pre-specified set of "true" parameters, then stochastic approximation was used to recover these parameters from the simulated presence-absence data.

3.2.2 Conditioning the network on the abiotic environment

The Markov networks fit in Harris (2015c) treated the α_i terms from Equations 1 and 2 as constants, meaning that the models expected each species to have the same conditional

occurrence probability in any location with a given set of facilitators and competitors. In this paper, I extend the model so that the local abiotic conditions can also affect species occurrence probabilities directly. Specifically, the α_i terms in these analyses are linear combinations of the local environmental conditions, represented by x_1 through x_5. By estimating the coefficients associated with each of these x variables for each species, we can account for their responses to the abiotic environment as well as to one another. In other fields, a model like this one, where a Markov network's parameters depend on external factors is called a conditional random field (Lee and Hastie 2012); these models have been used in a variety of contexts, from language models to image processing (Murphy 2012).

3.2.3   Coefficient estimation with stochastic approximation

A Markov network describes a probability distribution over possible assemblages. Since the model is a member of the exponential family it can be summarized without loss of information by its "sufficient statistics". For the model presented here, the sufficient statistics include: the number of occurrences for each species, the number of co-occurrences between each species, and the cross-product between the species and environment matrices (Azaele et al. 2010, Lee and Hastie 2012, Murphy 2012). Finding the maximum likelihood estimate for the model parameters is equivalent to minimizing the discrepancy between the values of the data's sufficient statistics and the expected sufficient statistics under the model (Bickel and Doksum 1977). Because fully-observed Markov networks, such as the ones analyzed here, have unimodal likelihood functions (Murphy 2012), it is possible to find the global optimum by iteratively reducing the discrepancies between the sufficient statistics of

the model and of the data to zero.

In order to reduce the discrepancy between the observed and predicted sufficient statistics, the `rosalia` calculates each value exactly, averaging over all possible presence-absence combinations. Stochastic approximation (Robbins and Monro 1951, Salakhutdinov and Hinton 2012) instead estimates the expected values of the sufficient statistics by averaging over a more manageable number of simulated assemblages during each model-fitting iteration, while still retaining maximum likelihood convergence guarantees. The procedure iterates through the following three steps as many times as needed (50,000 for these analyses; see Appendix for annotated R code):

1. Simulate a set of assemblages from the current model parameters and calculate sufficient statistics for the sample.

2. Subtract the simulated sufficient statistics from the observed ones to calculate the approximate likelihood gradient

3. Adjust the model parameters to climb the approximate gradient, using a schedule of step sizes that satisfies the criteria in Chapter 6 of Powell (2007).

Here, the simulations in Step 1 used Gibbs sampling to generate examples of landscapes based on the model's current parameter estimates. While the simulated landscapes produced by Gibbs sampling are serially autocorrelated, statisticians have shown that this merely slows convergence to the maximum likelihood estimate rather than preventing it altogether (Younes 1999, Salakhutdinov and Hinton 2012).

The approximate likelihood gradients in Step 2 match the ones from Harris (2015c), except that they are averaged over a set of Monte Carlo samples rather than over all possible presence-absence combinations. These gradients were augmented with a momentum term (Hinton 2012) and by regularizers based on a logistic prior with location 0 and scale 2.0 (for environmental responses) or 0.5 (for pairwise interactions). The step size parameter in Step 3, α_t, decreased after each iteration according to a generalized harmonic sequence, $\alpha_t = 1000/(999 + t)$, which satisfies the criteria in Powell (2007).

3.2.4 Simulated landscapes with known interactions

In order to assess the models' ability to recover the "true" parameters that generated a presence-absence matrix of interest, I first needed to generate such matrices from known processes.

To demonstrate that my stochastic approximation implementation could converge to the global optimum, I first simulated a landscape with 20 species and 500 sites, using Gibbs sampling. These landscapes had few enough species that the `rosalia` package for estimating Markov networks (Harris 2015a) could find the penalized maximum likelihood estimates for the Markov network in a reasonable amount of time. For each of these simulated landscapes, I then fit the same Markov network model to the simulated data using the exact approach from Harris (2015c) and the stochastic approximation method described above.

I then simulated a large landscape with 250 species and 2500 sites, representing a large data set roughly the size of the North American Breeding Bird Survey. Each site on the landscape was represented by 5 environmental variables, which were drawn independently

51

from Gaussian distributions with mean 0 and standard deviation 2.5.

Each species was assigned two sets of coefficients. The coefficients determining species'
responses to the environment were each drawn from standard normal distributions. The
coefficients describing species' pairwise interactions were drawn from a mixture of normals
(see Appendix) so that most interactions were weak and negative, but a few strong positive
and interactions also occurred.

The coefficients described above are sufficient to define each species' conditional
occurrence probability (i.e. its probability of occurrence against any given backdrop of biotic
and abiotic features). I used these conditional probabilities to produce examples of
communities that were consistent with the competition parameters and with the local
environmental variables via Markov chain Monte Carlo (specifically, 1000 rounds of Gibbs
sampling; see code in the Appendix). I then used the methods from the next section to
attempt to infer the underlying parameters from the simulated data.

3.3   Results

For the smaller communities, the squared deviations between the exact estimates produced
by the `rosalia` package and the ones produced by stochastic approximation quickly decayed
to negligible levels (Figure 2A), indicating that the stochastic approximation procedure was
implemented correctly and worked as the mathematical theory predicts.

For the larger landscape, I found that the stochastic approximation approach achieved
reasonably good performance after ten minutes of optimization (Figure 2B). After 50,000
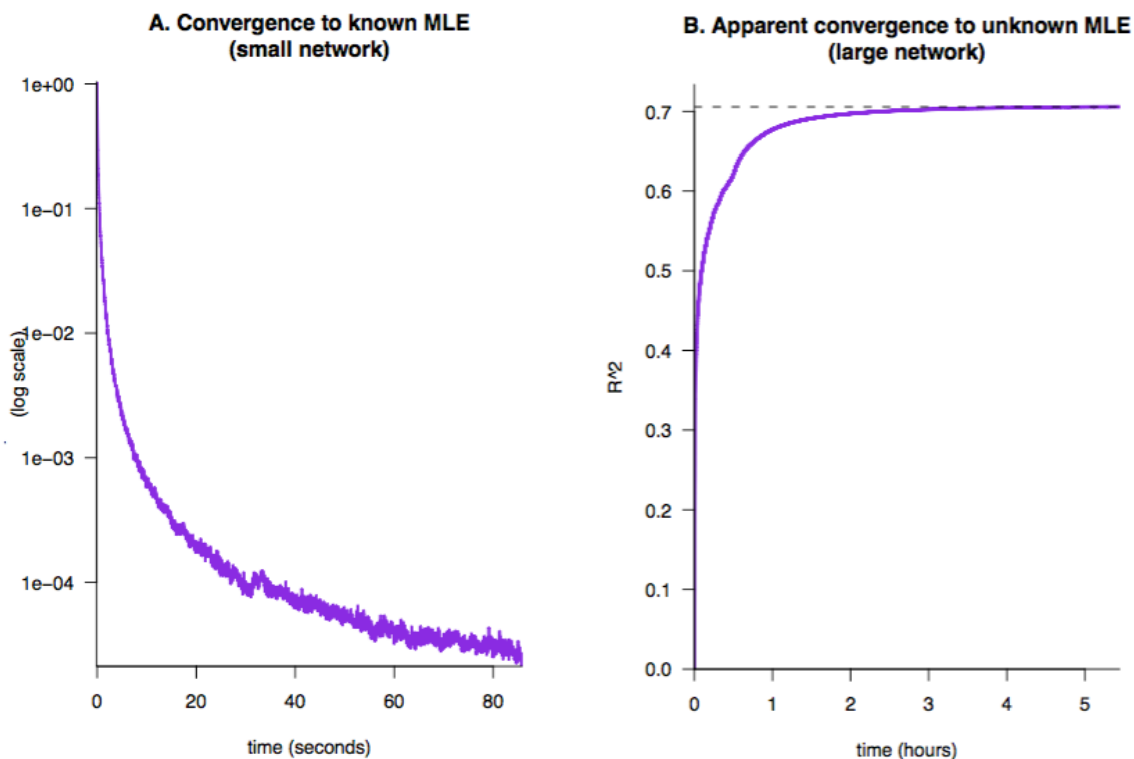
Figure 3.2: Stochastic approximation improves the model over time. **A.** With small networks where the exact maximum likelihood estimates are known, stochastic approximation approaches the global optimum very quickly (note the log scale of the y axis). Within a few seconds, the mean squared deviation between the approximate and exact estimates drops to less than 0.01; in less than a minute, it drops below 0.0001. For comparison, the `rosalia` package took about six minutes to find the maximum likelihood estimate. **B.** With larger networks, where the maximum likelihood estimate cannot be calculated exactly, stochastic approximation converges to an apparent optimum that explains most (but not all) of the variation in the "true" parameter values.
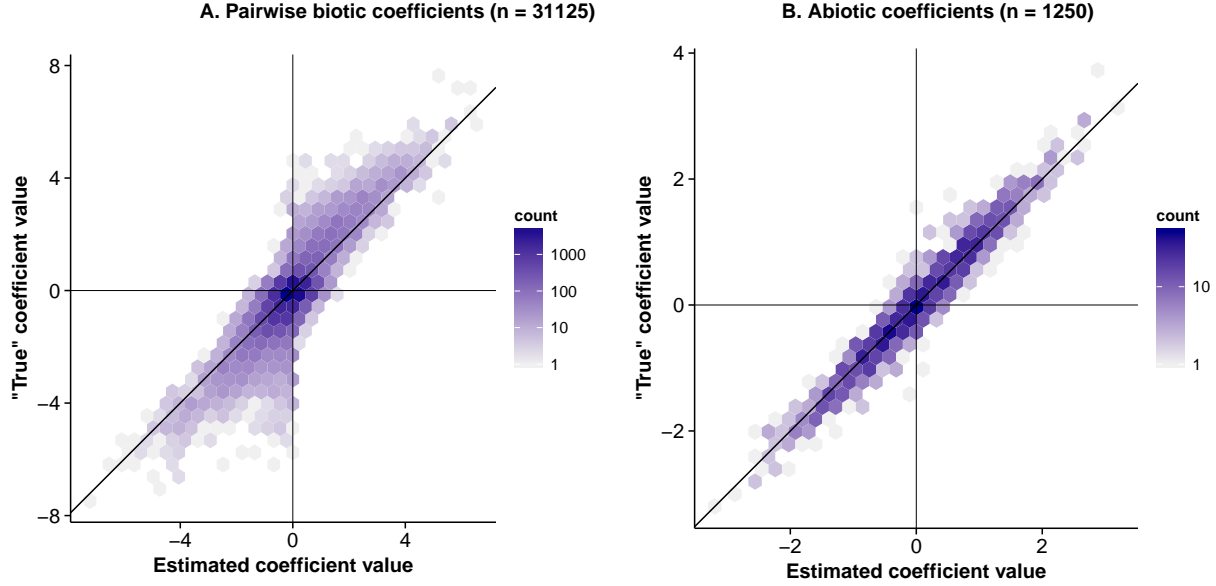
Figure 3.3: "True" versus estimated network parameters for all 250 species' responses to the other 249 species in the data set (**A.** $R^2 = .71$) and to the 5 abiotic variables (**B.** $R^2 = .95$).

iterations (about 5.4 hours on my laptop), the model was able to recover more than two thirds of the variance in species' pairwise interactions($R^2 = 0.71$; Figure 3A), and nearly all of the variance in their responses to environmental variables ($R^2 = 0.95$; Figure 3B).

3.4   Discussion

For decades, ecologists have relied on poor test statistics for inferring species interactions from observational data (Harris 2015c). As shown here and in Harris (2015c), however, these inferences can only be made reliably by methods that control for other species in the network, such as partial correlations or Markov networks. The biggest computational problem with large Markov networks is the impossibility of averaging over all their possible states, but these results demonstrate that this step can be avoided in ecological analyses. Stochastic approximation is able to recover the same estimates as exact methods for smaller

networks, while scaling gracefully to networks with hundreds of species.

The largest downside of using stochastic approximation instead of the exact methods introduced in Harris (2015c) is the difficulty in generating confidence intervals. The `rosalia` package can produce confidence intervals based on the Hessian matrix, but it is not feasible to calculate this matrix for large networks. It may turn out that the best way to generate confidence intervals for large networks is by repeatedly fitting the model to bootstrapped samples of the data. Alternatively, ecologists may be able to take advantage of the advanced Monte Carlo techniques introduced in Murray et al. (2012) to sample from the posterior distribution over possible parameter values, which could be used to generate credible intervals for the parameter estimates. Other advanced Monte Carlo methods (e.g. Salakhutdinov 2009) may also speed up convergence to the maximum likelihood estimate in cases where serial autocorrelation in the Gibbs sampler is too strong for effectively sampling the space of possible landscapes.

As the number of potentially-interacting species increases, the number of adjustable parameters increases even faster, so overfitting becomes a major concern. Fortunately, a number of good regularizers have been proposed. Of these, some of the most interesting options include L_1 regularization, which ensures that many of the estimated coefficients are exactly zero (Lee and Hastie 2012). This sparsity matches ecologists' intuition that many species from different guilds are unlikely to interact much at all (Faisal et al. 2010), and also produces computational benefits for model estimation. Of course, the best regularizers will take advantage of specific ecological knowledge (e.g. from field experiments, natural history, or trait data) to provide information about individual pairwise interactions, rather than

about their overall distribution.

These richer sources of information will be especially important for cases where ecologists expect that the interactions between species are asymmetric (e.g. where only one species is affected by an interaction or where one species benefits at a cost to the other). With snapshot observations of species' spatial associations, as discussed here, interactions must be reduced to a single number describing the net association, as in undirected models such as the Markov networks presented here (Schmidt and Murphy 2012). With richer data types, however, ecologists could more easily learn about asymmetric interactions such as facilitation, predation, and parasitism.

Even without better data, ecologists have a number of options for expanding beyond simple Markov networks in a number of ways that would improve their ability to address a wider range of questions. This paper demonstrated that it is possible to simultaneously estimate species' responses to the abiotic and environment and to one another, but many other extensions are possible. For example, ecologists could condition the model on variables whose values have not been measured (e.g. partially-observed Markov networks, or the approximate networks in the `mistnet` package; cf. Pollock et al. (2014)). This would allow ecologists to account for measurement error and for other ecologically-important factors that can be difficult to measure. Ecologists should also explore higher-order networks, where one species' presence can affect the relationship between two other species (Whittam and Siegel-Causey 1981, Tjelmeland and Besag 1998).

A Appendices to Chapter 1

A.1 Data

Code for this section can be found in

`mistnet/extras/BBS-analysis/data_extraction/data-extraction.R` in the version

control repository at https://github.com/davharris/mistnet/.

A.1.1 Partitioning training and test data

A set of evenly-spaced coordinates were selected across the Earth's surface using

`regularCoordinates(12)` in the `geosphere` R package (Hijmans et al. 2012). The 1559

routes that were more than 300 km from all of these coordinates were included in the

training set, while the 280 routes that were less than 150 km away from one of them were

included in the test set.

A.1.2 Criteria for including a species in the data set

Hybrids and other ambiguous taxa were identified using simple heuristics (e.g. if the listed

common or Latin name included the word "or"). Such species were excluded from the

analysis; code for the heuristics used is available on the mistnet package's version control

repository (`mistnet/extras/BBS-analysis/data_extraction/species-handling.R`). I

also omitted any species that were observed along fewer than 10 training routes to ensure

that enough data points were available for cross-validation. This left a pool of 368 species for

analysis.

A.1.3 Included climate predictors

The following eight climate predictors were selected out of the 19 Bioclim variables in the Worldclim data set, omitting nearly-collinear variables identified with the `findCorrelation` function in the `caret` package (Kuhn et al. 2012):

bio2: Mean Diurnal Range bio3: Isothermality bio5: Max Temperature of Warmest Month bio8: Mean Temperature of Wettest Quarter bio9: Mean Temperature of Driest Quarter bio15: Precipitation Seasonality bio16: Precipitation of Wettest Quarter bio18: Precipitation of Warmest Quarter

A.2 Neural network structure

A feed-forward neural network, such as a mistnet model, performs a series of simple mathematical operations in nodes, called neurons, which are organized into layers. Each neuron receives a vector of inputs from the previous layer, and ultimately from the environmental variables (both measured and simulated). The neuron's response to each input variable is governed by a coefficient ("weight"), such that the neuron's total activation equals the dot product of the input vector with its weight vector, plus an intercept term called a bias. The neuron may optionally perform a nonlinear operation on this activation level before passing it on as input to the following layer.

In the network presented here:

- The neurons in the first hidden layer perform output a piecewise linear function of their inputs, $f(x) = \max(0, x)$ (Zeiler et al. 2013).

- The neurons in the second hidden layer simply return their own inputs: $f(x) = x$. This layer functions solely to reduce the dimensionality of the network, as described in the main text.

- The output layer predicts occurrence probabilities for each species, with each species' neuron transforming its inputs to a probability between 0 and 1 with a logistic ("sigmoid") nonlinearity, $f(x) = e^x/(1 + e^x)$.

In a stochastic neural network (Neal 1992, Tang and Salakhutdinov 2013) such as mistnet, neurons can respond both to observed and simulated inputs. The simulated inputs represent samples from a probability distribution over possible unobserved environmental variables. Depending on the values taken by these unobserved (latent) variables, different values will be passed forward through the network and different probability distributions will be produced over possible species assemblages. This enables the network to represent the possibility that two locations with similar environmental data will have systematically different species composition, owing to unmeasured variation. By default, the latent variables are simulated from standard normal distributions.

A.3 Model fitting

If the neurons' nonlinear activation functions are differentiable almost everywhere, then one can use the chain rule from calculus to find the gradient of the model's log-likelihood with respect to every coefficient. The associated algorithm, called backpropagation, allows for local optimization by climbing this gradient (Murphy 2012).

Mistnet models' predictions, and thus their error gradients, depend on unobserved environmental variation, which means that they cannot be calculated exactly without solving difficult integrals. However, one can collect Monte Carlo samples of possible error gradients, since we have prior distributions over these unobserved variables (by default, these prior distributions are all standard Gaussians). Given the observed assemblage, one can therefore use an importance sampler to average over possible likelihood gradients, yielding an estimated gradient (Tang and Salakhutdinov 2013). The mistnet code then adjusts the model parameters in the direction of this estimate.

Tang and Salakhutdinov (2013) show that this procedure is an example of generalized expectation maximization (Neal and Hinton 1998), which means that it will increase a lower bound on the model's log-likelihood until that bound coincides with a local maximimum.

A.4   Model configuration

A.4.1   BRT

All these analyses used the `gbm` package. For each species, I evaluated BRT models with `interaction.depth` of 2, 5, and 8, and with up to 10,000 trees, using the default learning rate of 0.001. The number and depth of the trees was chosen by separate 5-fold cross-validations for each species (Murphy 2012). See `mistnet/extras/BBS-analysis/BBS_evaluation/gbm.R`.

A.4.2 Deterministic neural net

`nnet`'s hyperparameters were optimized using random search (Bergstra and Bengio 2012). During this search, the number of hidden units was sampled uniformly between 1 and 50 and the weight decay was sampled from an exponential distribution with rate parameter 1. The model was allowed 1,000 BFGS iterations to optimize the log-likelihood in each configuration. The model was fit with 8 different configurations of hidden layer sizes and weight decay values, and the best configuration was selected by five-fold cross-validation. See `mistnet/extras/BBS-analysis/BBS_evaluation/nnet-evaluation.R`.

A.4.3 BayesComm

I used a development version of BayesComm. This version can be downloaded and installed using the `devtools` package using the following command:

`install_github("goldingn/BayesComm", ref =`

`"0d710cda46a6e7427a560ee5b816c8ef5cd03eeb").`

I used the "full" model type, which models species' responses to both observed and latent environmental factors. BayesComm performed 42,000 rounds of Gibbs sampling, discarding the first 2,000 values as "burn-in" and retaining every 80th sample thereafter. This left 500 samples, which was a small enough number to fit in 8 gigabytes of memory with some room to spare for additional computations. See `mistnet/extras/BBS-analysis/BBS_evaluation/bayescomm.R`

A.4.4 mistnet

Worldclim variables were standardized to have zero mean and unit variance. The coefficients in each layer were initialized as random samples from a zero-mean Gaussian with a variance of 0.01. Bias (intercept) terms in the first layer were manually initialized at 1; bias terms in the second layer were left initialized at 0; bias terms in the final layer were automatically initialized as $\log(\frac{p}{1-p})$, where p is the proportion of routes where the corresponding species was observed. The variances of the Gaussian priors for the model's coefficients were adjusted in an empirical-Bayesian fashion every 10 iterations, using each prior's `update` method, with a minimum variance of 0.001. The means of the third layer's priors were also adjusted every 10 iterations, using the same function.

I tried 10 different hyperparameter configurations, chosen using random search (Bergstra and Bengio 2012). During this search, I varied the following hyperparameters:

- the number of routes to include in each round of gradient descent ("n.minibatch"), was sampled log-uniformly between 10 and 100. This range is commonly suggested in the neural network literature.

- the number of latent Gaussian variables ("sampler.size") was sampled log-uniformly between 5 and 20. This range was chosen based on my expectations about the likely number of uncorrelated environmental factors that would account for most of the variation in species composition.

- the number of importance samples to collect during each stage of gradient descent ("n.importance.samples") was sampled log-uniformly between 20 and 50, based on

Tang and Salakhutdinov (2013)'s suggestion of a range between 20 and 30, combined with some previous experience suggesting that larger values might work better than smaller ones.

- The number of nodes to include in the first hidden layer ("n.layer1") was sampled uniformly between 20 and 50, based on previous experience with neural network models fit to this data set.

- The number of nodes to include in the second hidden layer ("n.layer2") was sampled uniformly between 5 and 20, based on my expectations about the dimensionality of species' environmental responses.

- Coefficient updates were performed using adagrad (Duchi et al. 2011) and a learning rate of 0.1. Adagrad was chosen because of its relative insensitivity to hyperparameter values. The learning rate was chosen because it was the largest power of ten that did not introduce numerical problems.

These ranges will likely make for good starting points for many JSDM applications, with smaller values for the layer sizes and learning rates being more likely to perform well on smaller data sets. See also the `hyperparameters` vignette in the mistnet package, e.g. via `browseVignettes("mistnet")`.

For each hyperparameter configuration, I performed five-fold cross-validation, with 20 minutes allocated per fold for training. Cross-validation results are reported below, sorted by average route-level log-likelihood. The highest-performing set of hyperparameters (top row) was used to build the model used in the final analyses.

See

	n.minibatch	sampler.size	n.importance.samples	n.layer1	n.layer2	log.likelihood
1	79	10	28	37	15	-52.20
2	23	6	24	35	12	-52.26
3	15	15	25	43	10	-52.35
4	11	15	27	29	13	-52.37
5	37	13	36	31	15	-52.40
6	42	8	45	39	14	-52.57
7	81	8	22	23	11	-52.66
8	18	6	47	31	16	-53.05
9	46	20	28	22	9	-53.05
10	88	14	20	42	5	-55.29

`mistnet/extras/BBS-analysis/BBS_evaluation/mistnet_cross-validation.R` for the

code used in hyperparameter selection.

A.5 Variance decomposition

The variance decomposition at the beginning of the Results section was performed as follows.

For each species, I found the total variance in model predictions across both routes and

Monte Carlo samples, as well as the residual variance within routes (i.e. after climate was

accounted for). The proportion of non-climate variance for that species was calculated as

residual variance divided by total variance.

B Appendices to Chapter 3

B.1 Simulation and evaluation for the large landscape

```
library(rosalia)
library(mistnet)

# convenience function for adding intercepts to each column
`%plus%` = mistnet:::`%plus%`
logistic = binomial()$linkinv # logistic inverse link
```

```r
# Random bernoulli trial
rbern = function(p){rbinom(length(p), size = 1, prob = p)}

set.seed(1)
n_spp = 250      # number of species
n_loc = 2500     # number of locations
n_env = 5        # number of environmental predictors
n_gibbs = 5000   # number of Gibbs sampling iterations


# What portion of the coefficients should come from each
# mixture component?
type_frequencies = rmultinom(1, size = choose(n_spp, 2),
                            prob = c(.2, .5, .3))


# Vector of "true" interaction strengths.
# Three mixture components, shuffled by `sample`.
true_beta_vec = sample(
  c(
    rnorm(type_frequencies[[1]], 0, 2),
    rnorm(type_frequencies[[2]], 0, .2),
    rnorm(type_frequencies[[3]], -.5, .5)
  )
)

# Visualize the distribution of interaction terms
plot(density(true_beta_vec))

# Make a symmetric matrix of "true" beta coefficients
true_beta = matrix(0, nrow = n_spp, ncol = n_spp)
true_beta[upper.tri(true_beta)] = true_beta_vec
true_beta = true_beta + t(true_beta)

# Create a set of environmental predictors
true_env = matrix(rnorm(n_loc * n_env, 0, 2.5), nrow = n_loc)

# "true" responses of each species to each environmental variable
true_alpha_env = matrix(rnorm(n_spp * n_env, 0, 1), nrow = n_env)

# "true" intercepts for each species
true_alpha_species = rnorm(n_spp, 2)

# site-level intercept depends on species baselines
```

```r
# environmental responses
true_alpha = true_env %*% true_alpha_env %plus% true_alpha_species


# Initialize the `y` matrix.
# This will hold the "observed" presence-absence data
y = matrix(0.5, nrow = n_loc, ncol = n_spp)

# For each round of Gibbs sampling...
for(i in 1:n_gibbs){
  # For each species (in random order)...
  for(j in sample.int(n_spp)){
    # update its occurrence with samples from the
    # conditional distribution
    y[,j] = rbern(logistic(true_alpha[ , j] + y %*% true_beta[ , j]))
  }
}


# Calculate sufficient statistics of the data
y_stats = crossprod(y)
y_env_stats = t(true_env) %*% y


# Initialize the simulated landscape for stochastic approximation
y_sim = matrix(0.5, nrow = nrow(y), ncol = ncol(y))

# In this example, the true state of the environment is
# known without error
env = true_env

# Initialize species' responses to environment at 0.
# Also initialize the delta (change in parameter values from the
# previous optimization iteration) to zero, since no optimization
# has occurred yet
alpha_env = delta_alpha_env = matrix(0, nrow = n_env, ncol = n_spp)

# Initialize species' intercepts to match observed occurrence rates
# plus a small amount of regularization
alpha_species = qlogis((colSums(y) + 1) / (nrow(y) + 2))

# Initialize the deltas for the intercepts to zero
delta_alpha_species = rep(0, n_spp)

# Initialize pairwise interactions and deltas to zero
beta = delta_beta = matrix(0, nrow = n_spp, ncol = n_spp)
```

```r
# overall alpha depends on alpha_species and alpha_env.
# Will be filled in later, so can initialize it with zeros
# no delta alpha to initialize b/c alpha not optimized directly
alpha = matrix(0, nrow = n_spp, ncol = n_spp)


# Very weak priors on alpha terms, somewhat stronger on beta terms
alpha_env_prior = rosalia::make_logistic_prior(scale = 2)$log_grad
alpha_species_prior = rosalia::make_logistic_prior(scale = 2)$log_grad
beta_prior = rosalia::make_logistic_prior(scale = 0.5)$log_grad


initial_learning_rate = 1 # step size at start of optimization
maxit = 50000             # Number of rounds of optimization
start_time = as.integer(Sys.time())



# Record the R-squared values in this vector
r2s = numeric(maxit)

# Record the timing history in this vector
times = integer(maxit)

for(i in 1:maxit){
  ##############################
  # Gibbs sampling for predicted species composition
  ##############################

  # Update alpha
  alpha = env %*% alpha_env %plus% alpha_species

  # Sample entries in y_sim from their conditional
  # distribution (Gibbs sampling)
  for(j in sample.int(n_spp)){
    y_sim[,j] = rbern(logistic(alpha[ , j] + y_sim %*% beta[ , j]))
  }

  ##############################
  # Stochastic approximation for updating alpha and beta
  ##############################

  # Update learning rate and momentum
  learning_rate = initial_learning_rate * 1000 / (998 + 1 + i)
  momentum = .9 * (1 - 1/(.1 * i + 2))

  # Calculate sufficient statistics
```

```r
  y_sim_stats = crossprod(y_sim)
  y_sim_env_stats = t(env) %*% y_sim

  # Calculate the gradient with respect to alpha and beta.
  # Gradients are differences in sufficient statistics plus prior
  # gradients, all divided by the number of locations
  stats_difference = y_stats - y_sim_stats
  beta_grad = (stats_difference + beta_prior(beta)) / n_loc

  alpha_species_grad = (
    diag(stats_difference) +
      alpha_species_prior(alpha_species)
  ) / n_loc
  diag(beta_grad) = 0 # beta_ii is 0 by convention
  y_env_difference = y_env_stats - y_sim_env_stats
  alpha_env_grad = (y_env_difference +
                      alpha_env_prior(alpha_env))  / n_loc


  # Calculate parameter updates: gradient times learning rate
  # plus momentum times delta
  delta_beta = beta_grad * learning_rate + momentum * delta_beta
  delta_alpha_species = alpha_species_grad * learning_rate +
    momentum  * delta_alpha_species
  delta_alpha_env = alpha_env_grad * learning_rate +
    momentum  * delta_alpha_env

  # Add the deltas to the previous parameter values
  beta = beta + delta_beta
  alpha_species = alpha_species + delta_alpha_species
  alpha_env = alpha_env + delta_alpha_env

  # Record R-squared and timing
  r2s[i] = cor(
    true_beta[upper.tri(true_beta)],
    beta[upper.tri(beta)]
  )^2
  times[i] = as.integer(Sys.time()) - start_time
}

library(cowplot)
library(ggplot2)
out = plot_grid(
  ggplot(NULL, aes(x = beta[upper.tri(beta)],
                   y = true_beta[upper.tri(beta)])) +
```

```r
    stat_binhex() +
    xlab("Estimated coefficient value") +
    ylab("\"True\" coefficient value") +
    stat_hline(yintercept = 0, size = 1/8) +
    stat_vline(xintercept = 0, size = 1/8) +
    scale_fill_gradient(low = "#F0F0F0", high = "darkblue",
                        trans = "log10") +
    stat_abline(intercept = 0, slope = 1, size = 1/2) +
    coord_equal(),
  ggplot(NULL, aes(x = c(alpha_env), y = c(true_alpha_env))) +
    stat_binhex() +
    xlab("Estimated coefficient value") +
    ylab("\"True\" coefficient value") +
    stat_hline(yintercept = 0, size = 1/8) +
    stat_vline(xintercept = 0, size = 1/8) +
    scale_fill_gradient(low = "#F0F0F0", high = "darkblue",
                        trans = "log10") +
    stat_abline(intercept = 0, slope = 1, size = 1/2) +
    coord_equal(),
  nrow = 1,
  labels = c(
    "A. Pairwise biotic coefficients (n = 31125)",
    "B. Abiotic coefficients (n = 1250)"
  )
)
save_plot(
  "stochastic/estimates.pdf",
  out,
  base_aspect_ratio = 2,
  base_height = 6
)
```

B.2   Small landscape

```r
small_n_spp = 20
small_n_loc = 500
small_true_beta = true_beta[1:small_n_spp, 1:small_n_spp]
small_true_alpha = true_alpha_species - 2
small_y = matrix(0.5, nrow = small_n_loc, ncol = small_n_spp)

for(i in 1:n_gibbs){
  for(j in sample.int(small_n_spp)){
    small_y[,j] = rbern(logistic(small_true_alpha[j] +
                                 small_y %*% small_true_beta[ , j]))
```

```r
  }
}

exact_time = system.time({
  exact = rosalia(small_y,
                  prior = make_logistic_prior(scale = 1), maxit = 500)
})

mle_beta = exact$beta

maxit_small = 50000
start_time_small = as.numeric(Sys.time())

beta_prior_small = rosalia::make_logistic_prior(scale = 1)$log_grad
alpha_species_prior_small =
  rosalia::make_logistic_prior(scale = 1)$log_grad

alpha_small = qlogis((colSums(small_y) + 1) / (nrow(small_y) + 2))
delta_alpha_small = rep(0, small_n_spp)
beta_small = delta_beta_small =
  matrix(0, nrow = small_n_spp, ncol = small_n_spp)

y_sim_small = matrix(0.5, nrow = small_n_loc, ncol = small_n_spp)

y_stats = crossprod(small_y)

mses = numeric(maxit_small)
small_times = numeric(maxit_small)

for(i in 1:maxit_small){
  ##############################
  # Gibbs sampling for predicted species composition
  ##############################

  # Sample entries in y_sim from their conditional distribution
  # (Gibbs sampling)
  for(j in sample.int(small_n_spp)){
    y_sim_small[,j] = rbern(
      logistic(alpha_small[j] +
                 y_sim_small %*% beta_small[ , j])
    )
  }

  ##############################
```

```r
  # Stochastic approximation for updating alpha and beta
  ################################

  # Update learning rate and momentum
  learning_rate = initial_learning_rate * 1000 / (998 + 1 + i)
  momentum = .9 * (1 - 1/(.1 * i + 2))

  # Calculate sufficient statistics
  y_sim_stats = crossprod(y_sim_small)

  # Calculate the gradient with respect to alpha and beta
  stats_difference = y_stats - y_sim_stats
  beta_grad = (
    stats_difference + beta_prior_small(beta_small)
  ) / small_n_loc

  alpha_species_grad = (
    diag(stats_difference) +
      alpha_species_prior_small(alpha_small)
  ) / small_n_loc
  diag(beta_grad) = 0

  # Calculate parameter updates
  delta_beta = beta_grad * learning_rate +
    momentum * delta_beta_small
  delta_alpha_species = alpha_species_grad * learning_rate +
    momentum  * delta_alpha_small

  beta_small = beta_small + delta_beta
  alpha_small = alpha_small + delta_alpha_species

  mses[i] = mean((beta_small[upper.tri(beta_small)] -
                  exact$beta[upper.tri(exact$beta)])^2)
  small_times[i] = as.numeric(Sys.time()) - start_time_small
}
```

B.3 Plotting results

```r
pdf("stochastic/convergence.pdf", width = 10, height = 7)
par(mfrow = c(1, 2), las = 1)

plot(small_times, mses, type = "l",
     ylab = "mean square deviation from MLE\n(log scale)",
     xlab = "time (seconds)",
```

```r
      main = "A. Convergence to known MLE\n(small network)",
      yaxs = "i", bty = "l", log = "y", xaxs = "i", lwd = 2,
      col = "blueviolet", xlim = range(c(-.1, small_times)))

plot(c(0, times / 60 / 60), c(0, r2s), type = "l",
      xlab = "time (hours)", ylab = "R^2",
      ylim = c(0, max(r2s) * 1.04),
      main = "B. Apparent convergence to unknown MLE\n(large network)",
      yaxs = "i", xaxs = "i", lwd = 2,
      xlim = range(c(-.05, times / 60 / 60)),
      bty = "l", col = "blueviolet")
abline(h = max(r2s), col = "#00000080", lty = 2, lwd = 2)
dev.off()
```

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.

Bergstra, J., and Y. Bengio. 2012. Random search for hyper-parameter optimization. The Journal of Machine Learning Research 13:281–305.

Bickel, P., and K. Doksum. 1977. Mathematical Statistics: Basic Ideas and Selected Topics. San Francisco: Holden—Day.

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., and D. Simberloff. 1979. The assembly of species communities: Chance or competition? Ecology:1132–1140.

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.

Duchi, J., E. Hazan, and Y. Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. The Journal of Machine Learning Research 12:2121–2159.

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. 2015c. Estimating species interactions from observational data with Markov networks. bioRxiv.

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.

Hastings, A. 1987. Can competition be detected using species co-occurrence data? Ecology:117–123.

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.

Hijmans, R. J., E. Williams, and C. Vennes. 2012. Geosphere: Spherical Trigonometry.

Hinton, G. E. 2012. A practical guide to training restricted boltzmann machines. Pages 599–619 *in* Neural Networks: Tricks of the Trade. Springer.

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.

Kuhn, M., J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, and others. 2012. Caret: Classification and regression training. R package version 5:023.

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.

Lessard, J.-P., M. K. Borregaard, J. A. Fordyce, C. Rahbek, M. D. Weiser, R. R. Dunn, and N. J. Sanders. 2011. Strong influence of regional species pools on continent-wide structuring of local communities. Proceedings of the Royal Society of London B: Biological Sciences.

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.

Murray, I., Z. Ghahramani, and D. MacKay. 2012. MCMC for doubly-intractable distributions. arXiv preprint arXiv:1206.6848.

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.

Powell, W. B. 2007. Approximate Dynamic Programming: Solving the curses of dimensionality. John Wiley & Sons.

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.

Robbins, H., and S. Monro. 1951. A Stochastic Approximation Method. The Annals of Mathematical Statistics 22:400–407.

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.

Salakhutdinov, R. 2009. Learning in Markov random fields using tempered transitions. Advances in neural information processing systems 22:1598–1606.

Salakhutdinov, R., and G. Hinton. 2012. An efficient learning procedure for deep Boltzmann machines. Neural Computation 24:1967–2006.

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.

Younes, L. 1999. On the convergence of Markovian stochastic algorithms with rapidly decreasing ergodicity rates. Stochastics: An International Journal of Probability and Stochastic Processes 65:177–228.

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.

Zeiler, M. D., M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, and others. 2013. On rectified linear units for speech processing. Pages 3517–3521 *in* Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE.