APPENDIX A: MATHEMATICAL DETAILS OF RANDOM-EFFECTS ORDINATION

STEVEN C. WALKER AND DONALD A. JACKSON

Abstract. In this appendix, we provide the mathematical details of random effects ordination that were not explicitly described in the main text. We begin with a quick review of statistical modeling and information criteria. We then give a general treatment of random-effects ordination. Three other sections include details on the three specific approaches to random-effects ordination that are considered in the main text: factor analysis; PCA; latent trait models. We conclude with a small argument on why the CVIC (cross-validation information criterion) will have practically no bias.

Contents

I. Probabilistic statistical modeling background

Probabilistic statistical models of ecological study systems are extensively developed elsewhere (Hilborn and Mangel 1997; Burnham and Anderson 2002; Clark 2007; Bolker 2008), and so we will only briefly highlight the important concepts here. Statistical models assign a probability (or probability density for continuous data) to each possible data set, $\mathbf{Y}$, that could be observed from a study system. Relatively high probabilities indicate that $\mathbf{Y}$ is not a surprising observation, according to the model. These probabilities, $P(\mathbf{Y}|\boldsymbol{\theta})$ depend on a vector of parameters, $\boldsymbol{\theta}$; examples of parameters include means, intercepts, slopes, residual variances, carrying capacities, stage-specific mortalities, etc. Because we often lack theoretical considerations that would lead us to a choice for $\boldsymbol{\theta}$, we often use data to estimate it; such estimates, $\hat{\boldsymbol{\theta}}$, will be denoted with a hat. A common estimation procedure (called maximum likelihood) is to choose $\hat{\boldsymbol{\theta}}$ such that $P(\mathbf{Y}|\boldsymbol{\theta})$ is maximized at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, where $\mathbf{Y}$ is a data set that was actually observed. This statistical modeling framework provides a natural and general concept of goodness-of-fit: fitted models that assign high probability (density) to the observed data provide a better fit. It is more common to work on the log-scale because

of mathematical conveniences (Bolker 2008) and so,

$$\text{GOF} = \log(P(\mathbf{Y}|\hat{\boldsymbol{\theta}})) \tag{A.1}$$

is used as a general goodness-of-fit measure. Maximum likelihood estimation chooses parameters that maximize this GOF criterion.

Probabilistic prediction. It is one thing for a model to have high GOF, it is quite another for it to predict new data; prediction success is a stronger assessment of a model because fitting is easier than prediction. More parameters makes it easier for maximum likelihood to tailor the model to the observed data. Such over-fitting results in models that lack generality and therefore poorly predict new data from the modeled study system. Therefore, it is important to choose an appropriate level of complexity for our statistical models and we will show how to use information criteria to help make this choice in ordination analysis. One of the benefits of random-effects ordination is that it explicitly generates quantitative predictions about observational units not yet sampled. We first consider statistical prediction more generally before applying it to random-effects ordination.

It is easy to make a naïve prediction from a regression line. For a given value, x, of the independent variable, take the value, $a+bx$, as the prediction. Such a prediction is essentially certain to be wrong in the sense that the prediction will differ from the true value by some non-zero amount; in other words, there will be error in the prediction. However, there are ways to make predictions that take this inevitable error into account. For example, one type of probabilistic prediction involves the prediction intervals, ellipses, and regions that we considered in the main text (Fig. 1). A similar example with multivariate presence-absence data involves predicting the probability of co-occurrence of two or more species, as opposed to deterministically asserting co-occurrence or not.

Information criteria are based on another type of prediction, which generalizes the likelihood-based goodness-of-fit measure (Eq. A.1). Here we measure the success with which a model fitted to one data set, $\mathbf{Y}_\text{A}$, probabilistically predicts a new data set, $\mathbf{Y}_\text{B}$, by the log-probability density that the fitted model associates with $\mathbf{Y}_\text{B}$,

$$\log(P(\mathbf{Y}_\text{B}|\hat{\boldsymbol{\theta}}_\text{A})) \tag{A.2}$$

where $\hat{\boldsymbol{\theta}}_\text{A}$ is the vector of parameter estimates based solely on the first data set, $\mathbf{Y}_\text{A}$. Eq. A.2 assesses fitted models, $\hat{\boldsymbol{\theta}}_\text{A}$, by how likely they are to generate data that matches new data, $\mathbf{Y}_\text{B}$, from the same study system. The difference between this measure and the goodness-of-fit measure is that the data that were used to fit the model, $\mathbf{Y}_\text{A}$, are different from the data being predicted, $\mathbf{Y}_\text{B}$. Complex models are influenced by the idiosyncrasies of the data that happened to be sampled, making them poor at predicting new data. For this reason, the goodness-of-fit advantage enjoyed by complex models is of limited interest; over-fitted, complex models do not teach us much about our study system.

Expected Kullback-Leibler information. Prediction provides a tool for identifying when a model is too complex. This tool encourages us to prefer models that make better predictions of new data. The standard assessment of probabilistic predictions has become a quantity called the expected Kullback-Leibler information (Burnham and Anderson 2002), which is based on the measure of prediction success in Eq. A.2.

Expected Kullback-Leibler information is a theoretical quantity that considers a study design in which two data sets, $\mathbf{Y}_\text{A}$ and $\mathbf{Y}_\text{B}$, of equal sample size, n, are independently taken

from the same system. Then, ignoring an additive constant that depends on the study system, the expected Kullback-Leibler information is negative one times the average of the prediction success measure (Eq. A.2),

$$\mathrm{EKL} = -\mathrm{E}(\log(P(\mathbf{Y}_\mathrm{B}|\hat{\boldsymbol{\theta}}_\mathrm{A}))) \tag{A.3}$$

where E is the expectation (or average) with respect to repeated sampling of pairs of data sets from the study system.

Because of the negative sign in Eq. A.3, expected Kullback-Leibler information measures the expected error with which a model fitted to a data set of n observations predicts another data set of the same size. The fundamental principle underlying much of information theoretic model selection is that we should select models with low values of expected Kullback-Leibler information. Burnham and Anderson (2002) provided extensive arguments on why this fundamental principle is appropriate for statistical ecology.

There is a practical problem however; we usually only have a single data set. Yet calculating Eq. A.3 requires all possible data sets from the study system. This is not possible, especially since ecological study systems are not stationary and change over time and space. To solve this problem, information criteria are used to estimate expected Kullback-Leibler information from a single sample.

II. Information criteria

Akaike (1973) developed the first information criterion, AIC, which is used to estimate the expected Kullback-Leibler information of a maximum likelihood model fitting procedure,

$$\mathrm{AIC} = -2\log\left(P(\mathbf{Y}|\hat{\boldsymbol{\theta}})\right) + 2\mathcal{P} \tag{A.4}$$

where $\mathcal{P}$ is the number of free parameters fitted by maximum likelihood. The first term is negative two times the goodness-of-fit measure (Eq. A.1) and therefore measures lack-of-fit. The second term penalizes models in proportion to their complexity. If sample sizes are sufficiently large, $\frac{1}{2}$AIC is, up to an additive constant, an unbiased estimate of expected Kullback-Leibler information.

Cross-validation (Stone 1974) can be used to develop another information criterion. With the form of cross-validation used here, a parameter estimate, $\hat{\boldsymbol{\theta}}_{(i)}$, is made using all of the data except the ith observation, $\mathbf{y}_i$. Then the success of this fitted model at predicting the held-out observation is measured as $\log(P(\mathbf{y}_i|\hat{\boldsymbol{\theta}}_{(i)}))$, which is a special case of Eq. A.2. The important thing about this measure is that the data being predicted are not used in the fitting of the model that is being assessed. Subsequently, all of the other $n-1$ observations are left out, leading to the cross-validated measure of prediction success,

$$\mathrm{CV} = \sum_{i=1}^{n} \log(P(\mathbf{y}_i|\hat{\boldsymbol{\theta}}_{(i)})) \tag{A.5}$$

This equation is much like the goodness-of-fit measure (Eq. A.1) but without the circularity of predicting the same data that were used to fit the predictive model. The quantity $-$CV is also (up to an additive constant) an asymptotically unbiased estimate of expected Kullback-Leibler information (Yanagihara et al. 2006). Hence we define a cross-validation information criterion (CVIC) as,

$$\mathrm{CVIC} = -2\mathrm{CV} \tag{A.6}$$

Both AIC and CVIC are estimates of expected Kullback-Leibler information. Which criterion should be used? AIC is less variable (Efron 2004) and easier to calculate because it does not require repeated estimation. However, AIC can only be applied in certain circumstances whereas CVIC is much more general. For example, AIC can only be used with maximum likelihood estimation, whereas CVIC in principle may be applied to any estimation method. As we will see later, it is sometimes useful to consider methods of estimation other than maximum likelihood that lead to better predictive success and are therefore less prone to over-fitting. Hence we use both AIC and CVIC.

Despite this complementarity of AIC and CVIC, it is not appropriate to mix the two criteria in a single model-selection analysis. For example, it is not appropriate to calculate the AIC for one model and then the CVIC for another and then select the model with the smaller criterion. Because these two consistent estimates of expected Kullback-Leibler information have different properties, mixing them introduces the possibility that some models will gain an advantage over the others purely as a result of which criteria were used with which models. Hence if some models can only be assessed with CVIC—if they are not fitted by maximum likelihood for example—then all models must also be assessed with CVIC, despite the fact that AIC is less computationally demanding and less variable.

Although all of these measures for assessing probabilistic predictions and goodness-of-fit might seem overwhelming, it is important to remember that they are all just variations on the theme of $\log(P(\mathbf{Y}|\hat{\boldsymbol{\theta}}))$—the log-probability density of observed data under a fitted probability model. The concept of a probabilistic statistical model provides clear, general and defensible guidelines for assessing statistical analyses. This probabilistic approach encourages us to select models associated with low information criteria, because such models are expected to provide better probabilistic predictions of new data and hence are good candidates for working models of a particular study system.

III. General random-effects ordination model

Random-effects ordination models are specified in two steps. First we model the probability distribution, $P(\mathbf{x}_i)$, of the unobserved latent variables (i.e. the axes) for an observational unit, i. To keep our presentation simple, we assume that this distribution is fully specified and does not require any parameters although this restriction may be lifted. Second we model the probability distribution, $P(y_{ij}|\mathbf{x}_i, \boldsymbol{\theta}_j)$, of the observed data, y_{ij}, for the jth variable at the ith observational unit, given the axes, $\mathbf{x}_i$, and the parameters, $\boldsymbol{\theta}_j$, determining the relationship between the jth variable and the axes.

The first major assumption of our general model is that the mean, $\hat{y}_{ij}$, of the distribution, $P(y_{ij}|\mathbf{x}_i, \boldsymbol{\theta}_j)$, is given by,

$$g(\hat{y}_{ij}) = a(\boldsymbol{\theta}_j) + \sum_k f_k(\mathbf{x}_i, \boldsymbol{\theta}_j) \tag{A.7}$$

where g is a monotonically increasing 'link' function (e.g. identity link in factor analysis and the logit link in the latent trait model we present) and a and the f_k are functions. The first term, $a(\boldsymbol{\theta}_j)$, measures the overall central tendency of variable j. The terms in the summation are random effects. As discussed in the main text (lines): they are random because they depend on random ordination axes and they are effects because they measure deviations from an overall central tendency.

The second major assumption is that the variables are *conditionally* independent, given the value of the latent axes, $\mathbf{x}_i$; that is, for a given position, $\mathbf{x}_i$, in the ordination space, $\mathcal{O}$, the variables are independent. Therefore, we can write the conditional probability of all of the observed variables, $\mathbf{y}_i$, at the ith observational unit, as,

$$P(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta}) = \prod_j P(y_{ij}|\mathbf{x}_i, \boldsymbol{\theta}_j) \tag{A.8}$$

Given this conditional independence, how then can the model be used to explore associations between the variables? The key is to examine the model without explicit reference to the latent axes, which can be done by averaging over the conditional distributions associated with different positions in $\mathcal{O}$,

$$P(\mathbf{y}_i|\boldsymbol{\theta}) = \int_{\mathcal{O}} P(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta}) P(\mathbf{x}_i) d\mathbf{x}_i \tag{A.9}$$

This is the marginal distribution of the observed variables, and is central to random-effects modeling. Unlike the conditional distribution, this marginal distribution does not assume independence of the observed variables. Marginal dependencies arise out of conditional independence in the same way that we may observe a correlation between two variables that arises purely because they are both related to an unmeasured third variable (see Fig.1 in the main text for an example).

The probability of an entire data set, $\mathbf{Y}$, is,

$$P(\mathbf{Y}|\boldsymbol{\theta}) = \prod_i P(\mathbf{y}_i|\boldsymbol{\theta}) \tag{A.10}$$

Therefore, the log-likelihood is,

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_i \log(P(\mathbf{y}_i|\boldsymbol{\theta})) \tag{A.11}$$

Estimates, $\hat{\boldsymbol{\theta}}$, of the parameters, $\boldsymbol{\theta}$, can be obtained by maximizing this equation.

Once $\hat{\boldsymbol{\theta}}$ is obtained we can estimate the axis scores, conditionally on the observed data for each observational unit. By Bayes' theorem,

$$P(\mathbf{x}_i|\mathbf{y}_i, \hat{\boldsymbol{\theta}}) = \frac{P(\mathbf{y}_i|\mathbf{x}_i, \hat{\boldsymbol{\theta}}) P(\mathbf{x}_i)}{P(\mathbf{y}_i|\hat{\boldsymbol{\theta}})} = \frac{P(\mathbf{y}_i|\mathbf{x}_i, \hat{\boldsymbol{\theta}}) P(\mathbf{x}_i)}{\int P(\mathbf{y}_i|\mathbf{x}_i, \hat{\boldsymbol{\theta}}) P(\mathbf{x}_i) d\mathbf{x}_i} \tag{A.12}$$

A natural estimate of $\mathbf{x}_i$ is the mean of this distribution,

$$\hat{\mathbf{x}}_i = \int_{\mathcal{O}} \mathbf{x}_i P(\mathbf{x}_i|\mathbf{y}_i, \hat{\boldsymbol{\theta}}) d\mathbf{x}_i \tag{A.13}$$

This is the estimate that we employ here for both factor analysis and latent trait models in the main text. Although this method does use Bayes' theorem, those who have philosophical objections to Bayesian methods may accept this approach in good conscience. Most non-Bayesian statisticians do not object to such methods, largely because some non-Bayesian estimation method is first used to estimate the model parameters. There is a big philosophical difference between inferring latent variables and parameters with Bayes' theorem. For example, Bradley Efron, inventor of the bootstrap and vocal advocate for non-Bayesian statistics, had this to say while defending empirical Bayes methods: "Bayes' rule is satisfying,

convincing, and fun to use. But using Bayes' rule does not make one a Bayesian; always using it does, and that's where difficulties begin" (Efron 2003).

We also use the parameter estimates to make predictions. All of our predictions in this manuscript are based on the estimated marginal distribution, $P(\mathbf{y}_i|\hat{\boldsymbol{\theta}})$. The distribution of one set of variables, $\mathbf{y}_1$, given another set, $\mathbf{y}_2$, is the marginal, $P(\mathbf{y}|\hat{\boldsymbol{\theta}})$ (note that we have dropped the subscript i for simplicity of presentation), divided by the marginal distribution for the conditioning (i.e. predictor) variables,

$$P(\mathbf{y}_1|\mathbf{y}_2,\hat{\boldsymbol{\theta}}) = \frac{P(\mathbf{y}|\hat{\boldsymbol{\theta}})}{P(\mathbf{y}_2|\hat{\boldsymbol{\theta}})} = \frac{\int P(\mathbf{y}_1|\mathbf{x},\hat{\boldsymbol{\theta}})P(\mathbf{y}_2|\mathbf{x},\hat{\boldsymbol{\theta}})P(\mathbf{x})d\mathbf{x}}{\int P(\mathbf{y}_2|\mathbf{x},\hat{\boldsymbol{\theta}})P(\mathbf{x})d\mathbf{x}} \tag{A.14}$$

The mean of this distribution provides a regression equation when the variables are continuous, or conditional probabilities of occurrence when the variables are binary.

IV. Random-effects ordination with factor analysis

Factor rotations. In the main text we noted the fact that $\frac{1}{2}d(d-1)$ of the coefficients in $\mathbf{B}$ are not free parameters. The reason for this is that we can multiply the coefficient matrix by any orthogonal d-by-d matrix, $\mathbf{Q}$, and obtain a model with exactly the same likelihood (Johnson and Wichern 1992). To see this, note that from Eq. 4 in the main text the covariance matrix, $\widetilde{\mathbf{C}}$, for this transformed model is,

$$\widetilde{\mathbf{C}} = \mathbf{BQQ}'\mathbf{B}' + \boldsymbol{\Psi} \tag{A.15}$$

$$\widetilde{\mathbf{C}} = \mathbf{BIB}' + \boldsymbol{\Psi} \tag{A.16}$$

$$\widetilde{\mathbf{C}} = \mathbf{BB}' + \boldsymbol{\Psi} = \mathbf{C} \tag{A.17}$$

which is the same as the untransformed covariance matrix, $\mathbf{C}$, because $\mathbf{QQ}'$ equals the identity matrix, $\mathbf{I}$, as a result of the orthogonality of $\mathbf{Q}$. Therefore, maximum likelihood does not define a unique estimate of $\mathbf{B}$ but it does define a unique estimate of $\mathbf{C}$. The matrix $\mathbf{Q}$ is often called a rotation matrix. However, there is a non-geometric interpretation of $\mathbf{Q}$. Because we can choose any $\hat{\mathbf{B}}\mathbf{Q}$ as a maximum likelihood estimate, where $\hat{\mathbf{B}}$ is one ML estimate and $\mathbf{Q}$ is an arbitrary d-by-d orthogonal matrix, we may as well choose $\mathbf{Q}$ such that $\hat{\mathbf{B}}\mathbf{Q}$ is easy to interpret. There are a number of rotations that are used in the factor analysis literature. We use the varimax procedure (the default in `factanal`) for choosing $\mathbf{Q}$, because it tends to make the elements in the coefficient matrix either close to zero or large. Such a choice makes interpretation easier because we can unambiguously identify what correlations are being summarized by each axis. Although other procedures for choosing $\mathbf{Q}$ are available, a comparison of them is outside of our scope.

Full and null models. In the main text we compared factor analysis ordinations with a full and a null model. These models also follow a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$, but now $\mathbf{C}$ is no longer dependent on latent variables. However, the log-likelihood function,

$$\mathcal{L}(\boldsymbol{\mu},\mathbf{C}) = -\frac{n}{2}\left[p\log(2\pi) + \log|\mathbf{C}| + \mathrm{tr}(\mathbf{C}'\mathbf{S})\right] \tag{A.18}$$

still resembles the log-likelihood function for the factor analysis models. For the full model, $\mathbf{C}$ may be any valid covariance matrix—in technical terms $\mathbf{C}$ must be positive definite. For the null model, $\mathbf{C}$ is constrained to be diagonal—no non-zero correlations between variables.

For the full model, the maximum likelihood estimates of $\boldsymbol{\mu}$ and $\mathbf{C}$ are $\hat{\boldsymbol{\mu}} = \bar{\mathbf{y}}$ and $\hat{\mathbf{C}} = \hat{\mathbf{S}}$. The full model therefore has the intuitively appealing property that the maximum likelihood estimate of the mean and covariance matrix is simply the sample mean and sample covariance matrix. This result implies that latent variable models fit the data better when their estimates of $\boldsymbol{\mu}$ and $\mathbf{C}$ are more similar to the maximum likelihood estimates under the full model.

For the null model, the maximum likelihood estimates of $\boldsymbol{\mu}$ and $\mathbf{C}$ are the same as for the full model with the exception that all covariances (i.e. off-diagonal elements) in the sample covariance matrix are changed to zero; because the null model assumes that the dependent variables are uncorrelated, it does not use any sample information about correlations or covariances.

Regression with factor analysis. More insight into Eq. 13 in the main text can be obtained by rewriting it as,

$$\hat{\mathbf{y}}_1 = \hat{\mathbf{a}}_1 + \hat{\mathbf{B}}_1\hat{\mathbf{x}}_2 \tag{A.19}$$

where $\hat{\mathbf{x}}_2 = \hat{\mathbf{B}}_2'\hat{\mathbf{C}}_2^{-1}(\mathbf{y}_2 - \hat{\mathbf{a}}_2)$ is an estimate of the ordination axes using only the variables playing the role of predictors (compare with Eq. 12). Therefore, Eqs. 13 and A.19 can be seen as multiple multivariate linear regression equations. But these equations differ radically from standard regression, because the predictors are estimated axis scores, not correlated observed variables. If $\mathbf{y}_2$ were used as the predictors, as in least-squares regression, the resulting regression slopes will often be highly unstable due to multicollinearity. But the regression slopes in Eq. A.19 are stable because $\hat{\mathbf{x}}_2$ is estimated under the assumption that its components are un-correlated, thereby accounting for collinearity (Lawley and Maxwell 1973).

V. Random-effects ordination with PCA

From Eqs. 4 and 17 in the main text, the PPCAcor estimate of the correlation matrix is,

$$\mathbf{U}_d\mathbf{V}\mathbf{U}'_d + \hat{\psi}\mathbf{I} \tag{A.20}$$

where,

$$\mathbf{V} \equiv \boldsymbol{\Lambda}_d - \hat{\psi}\mathbf{I} \tag{A.21}$$

and $\boldsymbol{\Lambda}_d$ is a d-by-d diagonal matrix with the first d eigenvalues of the sample correlation matrix on the diagonal, $\hat{\psi}$ is the average of the $p - d$ smallest eigenvalues of the sample correlation matrix and $\mathbf{U}_d$ is a matrix with d columns given by the first d eigenvectors. The goal of this section is to assess how similar this PPCAcor estimate is to a true correlation matrix with block structure.

In the infinite sample limit, the sample correlation matrix is identical to the population correlation matrix. For our purposes here, the population correlation matrix has the block structure described in Section 9.3. This block structure allows us to say much more about the eigen-structure of the correlation matrix on which the PPCAcor estimates depend.

We are interested in PPCAcor for $d = g$ axes, because in the infinite sample limit each of the g groups should require only a single axis to summarize its covariance. By a result of (Rousson and Gasser 2004) (p. 545), the block structure of the correlation matrix ensures that the rows of $\mathbf{U}_g$ have at most a single non-zero element—each variable (row) is related to a single axis (column). Variables in the same group have their non-zero element in the

same column because they are associated with the same axis. But the Peres-Neto matrices have a special kind of block structure; correlations between pairs of variables in the same group are the same for all pairs in the group. For such matrices, by a result of (Johnson and Wichern 1992) (pp. 365-367), the non-zero elements in the mth column of $\mathbf{U}_g$ are all equal to,

$$u_m = \frac{1}{\sqrt{p_m}} \tag{A.22}$$

where p_m is the number of variables in the group. By the same result, the eigenvalue associated with the mth eigenvector is $1+(p_m-1)r_m$, where r_m is the within-group correlation for group m. Finally, combining the results of both Rousson and Gasser (2004) and Johnson and Wichern (1992) it follows that there are $p_m - 1$ residual eigenvalues equal to $1 - r_m$ for each group, m.

Given the above results about the eigen-structure of block matrices, we can deduce that $\hat{\psi}$ is,

$$\hat{\psi} = \frac{1}{p-g}\sum_{m=1}^{g}(p_m - 1)(1 - r_m). \tag{A.23}$$

We can also deduce that the diagonal elements of $\mathbf{V}$ consist of p_m repetitions of the same value, v_m, for each group and that,

$$\begin{aligned}
v_m &= 1 + (p_m - 1)r_m - \frac{1}{p-g}\sum_{m'=1}^{g}(p_{m'} - 1)(1 - r_{m'}) \\
v_m &= 1 + (p_m - 1)r_m - \frac{1}{p-g}\sum_{m'=1}^{g}(p_{m'} - 1) + \frac{1}{p-g}\sum_{m'=1}^{g}(p_{m'} - 1)r_{m'} \\
v_m &= 1 + (p_m - 1)r_m - 1 + \bar{r} \\
v_m &= (p_m - 1)r_m + \bar{r}
\end{aligned} \tag{A.24}$$

where, as in the main text,

$$\bar{r} = \frac{1}{p-g}\sum_{m=1}^{g}(p_m - 1)r_m. \tag{A.25}$$

Furthermore, the off-diagonal elements of $\mathbf{V}$ are all zero.

The PPCAcor estimated correlation, $\hat{\sigma}_m$, between two different variables of the same group, m, is the element in the matrix in Eq. A.20 that corresponds to the correlation between those two variables. By the rules of matrix algebra, this estimated correlation is,

$$\hat{\sigma}_m = u_m^2 v_m. \tag{A.26}$$

Substituting Eqs. A.22 and A.24 into this equation yields,

$$\hat{\sigma}_m = \left(1 - \frac{1}{p_m}\right) r_m + \left(\frac{1}{p_m}\right) \bar{r} \tag{A.27}$$

which is the required result (Eq. 18 in the main text).

For the correlation matrix PCA estimate presented by Johnson and Wichern (1992), the estimated correlation between variables i and j (both in group m) is given by,

$$\hat{\sigma}_m = \sum_{m'=1}^{g} \lambda_{m'} u_{im'} u_{jm'} \tag{A.28}$$

where λ_k is the kth eigenvalue and u_{jk} is the element of the eigenvector matrix associated with variable j and axis k. Then by the results of Rousson and Gasser (2004) and Johnson and Wichern (1992) reviewed above, only one of the terms in this sum is non-zero, corresponding to $m' = m$—this is because each variable is associated with a single axis only. Therefore,

$$\begin{aligned}
\hat{\sigma}_m &= \lambda_m u_{im} u_{jm} \\
\hat{\sigma}_m &= (1 + (p_m - 1) r_m) u_{im} u_{jm} \\
\hat{\sigma}_m &= (1 + (p_m - 1) r_m) u_m^2 \\
\hat{\sigma}_m &= (1 + (p_m - 1) r_m) \frac{1}{p_m} \\
\hat{\sigma}_m &= \left(1 - \frac{1}{p_m}\right) r_m + \left(\frac{1}{p_m}\right)
\end{aligned} \tag{A.29}$$

which is the required result (Eq. 21 in the main text).

VI. Random-effects ordination with latent trait models

This material describes the principles behind our `R` function, `ltm.ecol` (Supplement 2 and Appendix B). Our approach is based on a combination of the ideas in the following methodological papers: (Bock and Aitkin 1981; Woodruff and Hanson 1996; Rizopoulos 2006; Houseman et al. 2007; Friedman et al. 2010).

Approximating the ordination space discretely. Our approach begins with a discretization, $\mathcal{O}_g^2$, of the two-dimensional ordination space, which is a g-by-g regular lattice that is bounded between -3 and 3 along both axes (Fig. A1). The shading of each lattice point is proportional to the probability that a randomly sampled observational unit will be characterized by the corresponding point in $\mathcal{O}_g^2$. This ordination space is built by taking the Cartesian product,

$$\mathcal{O}_g^2 = \mathcal{O}_g \times \mathcal{O}_g \tag{A.30}$$

of one-dimensional ordination spaces, $\mathcal{O}_g = \{-3, -3 + \frac{6}{g-1}, -3 + \frac{12}{g-1}, ..., 3\}$, where $\mathcal{O}_g$ can be thought of as a single axis. Higher dimensional spaces can be produced by taking higher order Cartesian products of $\mathcal{O}_g$. This discretization idea for latent trait model fitting is clearly described by Woodruff and Hanson (1996).

Define the value of the ith site in $\mathcal{O}_g^2$ as the ordered pair (x_{i1}, x_{i2}) such that x_{i1} and x_{i2} are both in $\mathcal{O}_g$. The probability associated with each point, (x_1, x_2), in $\mathcal{O}_g^2$ is,

$$q_g(x_1, x_2) = \frac{\exp(-\frac{x_1^2 + x_2^2}{2})}{\sum_{(x_1, x_2) \in \mathcal{O}_g^2} \exp(-\frac{x_1^2 + x_2^2}{2})} \tag{A.31}$$

As $g \to \infty$, the lattice, $\mathcal{O}_g^2$, fills the space $[-3, 3] \times [-3, 3]$ and $q_g(x_1, x_2)$ goes to zero. Therefore, as with all continuous distributions, the probability of any single point in the space is zero in the limit of $g \to \infty$. However, the probabilities of sets of lattice points in

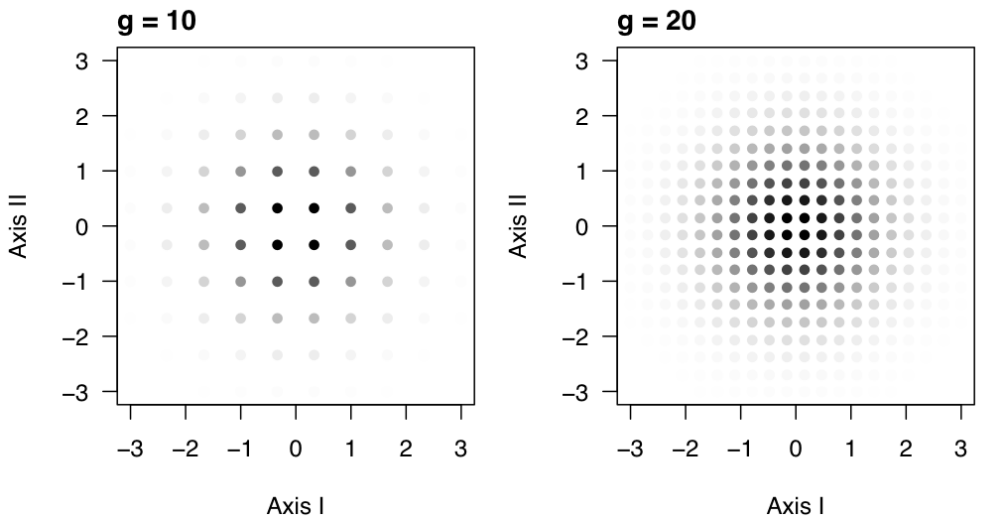

FIGURE A1. Ordination space approximation by a regular lattice.

particular regions will be non-zero as $g \to \infty$, if the regions stay fixed as g increases and have non-zero volume; in fact these probabilities will be equal to the probabilities given by a bivariate normal density, $\mathcal{N}_2(0, I)$, which is truncated outside of $[-3, 3] \times [-3, 3]$. Because most of the probability mass of $\mathcal{N}_2(0, I)$ is on $[-3, 3]\times[-3, 3]$, $q_g(x_1, x_2)$ approximates $\mathcal{N}_2(0, I)$ and the approximations get better as g increases. Therefore, $\mathcal{O}_g^2$ approximates the ordination space of the latent trait model given in the main text.

The major benefit of this approximation is that we can use the EM-algorithm for finite mixtures (Woodruff and Hanson 1996), which lets us avoid direct evaluation of extremely difficult integrals. In fact, our discretization can be considered the first step to a simple numerical integration scheme. There is a trade-off here, choosing g too large will make computation times prohibitively long but choosing g too small will make the approximation worse. In this paper, we used $g = 10$ which should be adequate for most ecological purposes. Interestingly, smaller g's makes the analysis more like a clustering approach whereas larger g's makes it more like ordination, because the space, $\mathcal{O}_g^2$, becomes 'more continuous'.

To put this lattice to work, we convert it into a matrix. Let $\mathbf{x} = [x_k]$ be the vector with the values in $\{-3, -3 + \frac{6}{g-1}, -3 + \frac{12}{g-1}, ..., 3\}$. Define the g^2-by-5 matrix,

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 & x_1 & x_1^2 \\ 1 & x_1 & x_1^2 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_1 & x_1^2 & x_g & x_g^2 \\ 1 & x_2 & x_2^2 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_2 & x_2^2 & x_g & x_g^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_g & x_g^2 & x_1 & x_1^2 \\ 1 & x_g & x_g^2 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_g & x_g^2 & x_g & x_g^2 \end{pmatrix} \tag{A.32}$$

The first column contains only 1's, and corresponds to the intercept. The second (fourth) and third (fifth) columns represent the first (second) axis and its square. Notice that there is one row in $\mathbf{X}$ for each point in the lattice, $\mathcal{O}_g^2$. Let $\mathbf{q}$ be a g^2-by-1 column vector with elements corresponding to the probabilities—given by the distribution, q_g, over $\mathcal{O}_g^2$—associated with each row in $\mathbf{X}$ (i.e. each point in $\mathcal{O}_g^2$).

Likelihoods and probabilties. Now that our ordination space and its probability distribution has been defined, we can compute all of the probabilities and expectations predicted from our model. In all of these expressions, the coefficients, b_{jk}, are analogous to elements of $\boldsymbol{\theta}$ in the general model (*nb.* in the main text we use b's and c's to distinguish linear and quadratic coefficients, but not here in the appendix). We begin with the conditional probability of observing species j at observational unit i given that the site is at the lth point in the lattice,

$$\log(p_{ij|l}) = \sum_k (y_{ij} b_{jk} x_{lk} - \log(1 + \exp(b_{jk} x_{lk}))) \tag{A.33}$$

The left hand side of this equation is analogous to $\log(P(y_{ij}|\mathbf{x}_i, \boldsymbol{\theta}_j))$ in the general model. The conditional probability of all variables (analogous to $P(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta})$; Eq.A.8), is given by,

$$\log(p_{i|l}) = \sum_j \log(p_{ij|l}) \tag{A.34}$$

The marginal distribution (i.e. $P(\mathbf{y}_i|\boldsymbol{\theta})$) is,

$$p_i = \sum_l p_{i|l} q_l \tag{A.35}$$

where q_l is the probability of being at lattice point l. Here is the first sign of a benefit of our lattice approximation. To see this, note that Eq.A.35 is a special case of Eq.A.9, where the integral has been replaced by a sum over the g^2 points in $\mathcal{O}_g^2$—so instead of computing an

extremely difficult integral we simply need to compute this sum. Using Bayes' theorem, we can also derive the conditional probability of being at lattice point l given the data at site i,

$$q_{l|i} = \frac{p_{i|l} q_l}{p_i} \tag{A.36}$$

This equation is used in the EM-algorithm for maximizing the log-likelihood function that we now describe.

We can now write down the log-likelihood for the entire observed data set, which is simply,

$$\mathcal{L}(\mathbf{B}) = \sum_i \log(p_i) \tag{A.37}$$

This is the likelihood equation that the `ltm` function attempts to maximizes. However, we found that maximizing this likelihood does not work very well for ecological data, because it leads to unrealistically certain estimated probabilities of occurrence. Therefore, we seek to maximize,

$$\Lambda(\mathbf{B}) = \mathcal{L}(\mathbf{B}) - \lambda \sum_j \sum_{k=2}^{5} |b_{jk}| \tag{A.38}$$

such that b_{j3} and b_{j5} are not positive for all species, which ensures hump- rather than S-shaped responses. The second term is a penalty for coefficients with large absolute values. This penalty becomes more severe as the regularization parameter, λ, is increased. Notice that we do not put a penalty on the intercepts, because there is a great deal of information about the part of the data modeled by the intercepts (i.e. the overall level of occurrence of each species). It is this equation that we seek to maximize using the EM-algorithm.

E-step. The EM-algorithm comes in two steps. The first step (called the E-step) is to calculate the so-called 'pseudo-data',

$$\begin{aligned} \nu_l &= \sum_i q_{l|i} \\ r_{jl} &= \sum_i y_{ij} q_{l|i} \end{aligned} \tag{A.39}$$

The first quantity, ν_l, is interpreted as an estimate of the number of sites in the sample that is located on the lth point in the lattice. The second quantity, r_{jl}, is interpreted as an estimte of the number of sites at lattice point l having species j present.

Non-regularized M-step. We first give an M-step for maximizing $\mathcal{L}(\mathbf{B})$ (Eq.A.37), before maximizing the regularized form (Eq.A.38). This M-step is not very efficient for maximizing $\mathcal{L}(\mathbf{B})$, but a simple modification yields a good algorithm for maximizing $\Lambda(\mathbf{B})$. This non-regularized M-step is based on the iteratively reweighted least-squares algorithm for computing logistic regression. Each M-step consists of one weighted least-squares problem for each coefficient, b_{jk}. The response variable of the least-squares problem for coefficient b_{jk} is (r_{jl}, ν_l) for $l = 1, ..., g^2$ and the predictor variable is the kth column, $\mathbf{x}^{(k)}$ of $\mathbf{X}$. There is also an offset term given by $\sum_{k' \neq k} b_{jk'} x_{lk'}$. The goal is to estimate b_{jk}, which is the only coefficient for species j that is excluded from the offset term—the other coefficients are held constant.

The probability of occurrence of species j at lattice point l is,

$$p_{j|l} = \frac{1}{1 + \exp(-\sum_{k'} b_{jk'} x_{lk'})} \tag{A.40}$$

The b_{jk} coefficients in this equation are obtained from the previous iteration. The weights for each lattice point are,

$$w_{jl} = \nu_l p_{j|l}(1 - p_{j|l}) \tag{A.41}$$

The iteratively reweighted least-squares algorithm makes use of the linearized logit transform,

$$\zeta_{jl} = b_{jk} x_{lk} + \frac{r_{jl} - \nu_l p_{j|l}}{w_{jl}} \tag{A.42}$$

Let $\boldsymbol{\zeta}_j = [\zeta_{jl}]$ be the vector containing the linearized transforms for species j. Then we may write the weighted least squares solution as,

$$b_{jk} = (\mathbf{X}^T \mathbf{W}_j \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}_j \boldsymbol{\zeta}_j \tag{A.43}$$

where $\mathbf{W}_j$ is a diagonal matrix with the weights for species j on the diagonal. We use this weighted-least squares estimator in the regularized M-step below.

Regularized M-step. This non-regularized M-step can be modified into a regularized M-step by passing each weighted least-squares solution through a soft-thresholding operator (Friedman et al. 2010). For the constant terms ($k = 1$),

$$b_{jk} \to b_{jk} \tag{A.44}$$

for the linear terms ($k = 2, 4$),

$$b_{jk} \to \begin{cases} b_{jk} - \lambda' & b_{jk} > 0, |b_{jk}| > \lambda' \\ b_{jk} + \lambda' & b_{jk} < 0, |b_{jk}| > \lambda' \\ 0 & |b_{jk}| < \lambda' \end{cases} \tag{A.45}$$

for the quadratic terms ($k = 3, 5$),

$$b_{jk} \to \begin{cases} b_{jk} + \lambda' & b_{jk} < 0 \\ 0 & \text{otherwise} \end{cases} \tag{A.46}$$

where,

$$\lambda' = \frac{\lambda}{\sum_l w_{jl} x_{kl}^2} \tag{A.47}$$

Full EM-algorithm. Algorithm 1 summarizes the main steps of our recommended approach to fitting presence-absence community data to latent trait models.

We choose an initial value for $\mathbf{B}$ using the first two axes of correspondence analysis. Let $\widetilde{\mathbf{X}}$ be an n-by-5 matrix with column one containing only ones, columns two and four containing the first and second correspondence analysis axes (centered to mean zero and scaled to variance one), and columns three and five containing their squares. Let $\widetilde{\mathbf{Y}}$ be the matrix that results from transforming each element, y_{ij}, of $\mathbf{Y}$, by $2y_{ij} - 1$. Then calculate the least-squares estimate, $\widetilde{\mathbf{B}}$, of the linear model with $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$ as predictor and response matrices. We then take as our initial estimate of $\mathbf{B}$,

$$\mathbf{B}_0 = \log(\frac{\widetilde{m}}{n - \widetilde{m}})\widetilde{\mathbf{B}} \tag{A.48}$$

Algorithm 1 Summary of the EM-algorithm for the ecological latent trait model

(0) Choose an initial value for $\mathbf{B}$.
(1) E-step:
  (a) Calculate each $\log(p_{ij|l})$ (Eq.A.33) for all i, j, l.
  (b) Sum over the species to obtain $\log(p_{i|l})$ for each i, l (Eq.A.34).
  (c) Calculate the marginal probabilities, p_i, for each site, i (Eq.A.35).
  (d) Use Bayes' theorem to convert these $p_{i|l}$ and p_i values into the $q_{l|i}$ for each i, l (Eq.A.36).
  (e) Use the $q_{l|i}$ to calculate the pseudo-data, ν_l and r_{jl} (Eq.A.39).
(2) M-step:
  (a) Loop over each type, $k = 1, ..., 5$, of coefficient.
    (i) Loop over each species, $j = 1, ..., m$.
      (A) Calculate $p_{j|l}$ (Eq.A.40) for each l using the current estimates of the coefficients in $\mathbf{B}$.
      (B) Calculate the weights, w_{jl} (Eq.A.41), for each l using these $p_{j|l}$ and the pseudodata from the E-step.
      (C) Calculate linearized logit transform, ζ_{jl}, for each l (Eq.A.42).
      (D) Update b_{jk} using Eq.A.43.
      (E) Pass this b_{jk} through the appropriate soft-thresholding operator (Eqs.A.44,A.45,A.46,A.47).
(3) Repeat 1-2 until convergence.

where,

$$\widetilde{m} = \min\{\max_j(n_j), n-1\} \tag{A.49}$$

and n_j is the number of observational units at which species j is present and n is the total number of observational units. We chose this method of selecting initial values because it has been convenient in practice. Some interesting work could be done in identifying better initial values. Our `ltm.ecol` function allows users to specify whatever initial values they would like.

We terminate the algorithm when the observed data log-likelihood—rounded to δ decimal places—does not change from one iteration to the next. When δ is smaller, convergence is judged more rapidly resulting in reduced accuracy but larger δ results in long computation times. We found that $\delta = 1$ produced satisfactory results with several data sets, in the sense that our qualitative interpretations were left unchanged by increasing δ.

In practice, we do not know what value of the regularization parameter, λ, will be best, and so we compare various models for a range of values, $\lambda_1 < \lambda_2 < ... < \lambda_\rho$. We begin with λ_1 and employ our EM-algorithm using the correspondence analysis method of choosing initial values. We then run the algorithm for λ_2 using the estimates for λ_1 as the initial values. We then continue in this fashion until λ_ρ. This approach is more computationally efficient than starting back at $\mathbf{B}_0$ for each value of λ, because the estimates for adjacent λ values will tend to be 'close' to each other—especially when the difference between λ values is small—resulting is faster convergence (Friedman et al. 2010). We choose λ by cross-validation and comparing the resulting CVIC values. Cross-validation is what really lengthens computing

time. When cross-validation computation time is too long, we recommend using $\lambda = 1$ as a reasonable default value.

Biplots and prediction. The estimate of the first and second ordination axes at observational unit, i, are,

$$\sum_l x_{l2} q_{l|i} \tag{A.50}$$

and

$$\sum_l x_{l4} q_{l|i} \tag{A.51}$$

This is a special case of Eq.A.13. We can also calculate the probability of observing species j at observational unit i given the pattern of all of the other species, $(-j)$, at i,

$$p_{ij|i(-j)} = \frac{\sum_l p_{i|l} q_l}{\sum_l p_{i|l}^{(-j)} q_l} \tag{A.52}$$

where $p_{i|l}^{(-j)}$ is defined by,

$$\log(p_{i|l}^{(-j)}) = \sum_k \sum_{j' \neq j} \log(p_{ij'|l}) \tag{A.53}$$

which is just Eq.A.34 with species j omitted. Various other probabilities can be calculated by leaving out various other species in the numerator and denominator of Eq.A.52.

VII. Why CVIC is practically unbiased

Let $\mathbf{Y}$ and $\mathbf{Z}$ be multivariate data sets of size n and $n-1$ respectively, such that all $2n-1$ multivariate observations have the same multivariate distribution. Let $\mathbf{y}_i$ be the ith observation in $\mathbf{Y}$ and $\mathbf{Y}_{(i)}$ be $\mathbf{Y}$ with the ith observation omitted. For convenience, I use a notation for probability densities that is slightly different from the main text. Let $P(\cdot|\mathbf{Z})$ indicate probability density given a model fitted to $\mathbf{Z}$. Then,

$$\text{CVIC} = -2 \sum_{i=1}^{n} \log(P(\mathbf{y}_i|\mathbf{Y}_{(i)})). \tag{A.54}$$

The expected value of CVIC is,

$$\begin{aligned} E(\text{CVIC}) &= E(-2 \sum_{i=1}^{n} \log(P(\mathbf{y}_i|\mathbf{Y}_{(i)}))) \\ &= -2 \sum_{i=1}^{n} E(\log(P(\mathbf{y}_i|\mathbf{Y}_{(i)}))) \\ &= -2 \sum_{i=1}^{n} E(\log(P(\mathbf{y}_i|\mathbf{Z}))) \\ &= -2nE(\log(P(\mathbf{y}_i|\mathbf{Z}))) \\ &= -2E(\log(P(\mathbf{Y}|\mathbf{Z}))). \end{aligned} \tag{A.55}$$

The first step from line one to line two follows by the linearity of expected value (Evans and Rosenthal 2002). The second step follows because $\mathbf{Z}$ and $\mathbf{Y}_{(i)}$ have identical distributions.

The third and fourth steps follow because all $\mathbf{y}_i$ have identical distributions. Note that the right-hand side of the final equation only differs from the expected Kullback-Leibler information because the training data, $\mathbf{Z}$, have sample size $n-1$ instead of n. Because the expected value of CVIC is equivalent to a quantity that is very similar to the expected Kullback-Leibler information, we do not expect the bias of CVIC to be substantial.

References

Akaike, H., 1973. Information theory as an extension of the maximum likelihood principle. In B. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest. Akademiai Kiado.

Bock, R. D. and M. Aitkin, 1981. Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika* **46**:443–459.

Bolker, B. M., 2008. Ecological Models and Data in R. Princeton University Press, Princeton, New Jersey.

Burnham, K. P. and D. R. Anderson, 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer, New York, second edition.

Clark, J. S., 2007. Models for Ecological Data. Princeton University Press, Princeton.

Efron, B., 2003. Bayesians, frequentists, and physicists. In *PHYSTAT2003*, pages 8–11, Stanford California. SLAC National Accelerator Laboratory.

Efron, B., 2004. The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association* **99**:619–642.

Evans, M. J. and J. S. Rosenthal, 2002. Probability and Statistics: The Science of Uncertainty. W.H. Freeman and Company, New York.

Friedman, J. H., T. Hastie, and R. Tibshirani, 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**:1–22.

Hilborn, R. and M. Mangel, 1997. The Ecological Detective. Princeton University Press, Princeton, New Jersey.

Houseman, E. A., C. Marsit, M. Karagas, and L. M. Ryan, 2007. Penalized item response theory models: Application to epigenetic alterations in bladder cancer. *Biometrics* **63**:1269–1277.

Johnson, R. A. and D. W. Wichern, 1992. Applied Mulitvariate Statistical Analysis. Prentice Hall, third edition.

Lawley, D. and A. Maxwell, 1973. Regression and factor analysis. *Biometrika* **60**:331–338.

Rizopoulos, D., 2006. ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software* **17**:1–25.

Rousson, V. and T. Gasser, 2004. Simple component analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **53**:539–555.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)* **36**:111–147.

Woodruff, D. J. and B. A. Hanson, 1996. Estimation of item response models using the EM algoirhtm for finite mixtures. Technical report, ACT.

Yanagihara, H., T. Tonda, and C. Matsumoto, 2006. Bias correction of cross-validation criterion based on Kullback-Leibler information under a general condition. *Journal of Multivariate Analysis* **97**:1965–1975.