Appendix S1
Why species tell more about traits than vice versa: Predictive analysis to connect species, traits, and environment

JAMES S. CLARK

In this supplement I provide mathematics for trait norms and co-dispersion, data used in the example, and modeling/computation background.

Trait norms and co-dispersion

To facilitate connections to joint trait distributions I provide norms and co-dispersion in matrix notation. In the following $t_{ms}$ is trait $m$ for species $s$, organized in a $M$ by $S$ matrix of traits by species matrix $T$. For a single trait the *trait norm* is

$$ \bar{u}_m = S^{-1} t'_m 1_S $$

where $t_m$ is the length-$S$ vector of trait values, the $m^{th}$ row of $T$, and $1_S$ is the length-$S$ vector of ones. For $M$ traits the trait norm (TN) is a vector

$$ TN = \bar{u} = S^{-1} T 1_S $$

This is the LQ matrix of RLQ analysis (Doledec et al. 1996, Brown et al. 2014, Dray et al. 2014).

The CWM value is a *weighted trait norm* taken for a vector of observed species weights at location $i$

$$ WTN_i = T w_i $$

Taken over a sample of $n$ observations the weighted trait norm is $WTN = n^{-1} T W 1_n$, where $W$ is the $n$ by $S$ species composition sample. $WTN$ varies between minimum and maximum trait values; these extremes obtain when the species having those values dominate a sample. Exploratory methods, including randomization techniques, are often used to evaluate $WTN$.

Although traits are fixed in CWM analysis they do have dispersion that can be represented in ways that permit analysis of species and environment effects. Trait co-dispersion independent of species diversity quantifies tendency for traits to be shared across species, the columns in trait-by-species matrix $T$,

$$ TC = S^{-1} (T - \bar{U})(T - \bar{U})' $$

where

$$ \bar{U} = \bar{u} 1'_s $$
is a $M$ by $S$ matrix, each column being the length-$M$ vector of trait norms $\mathbf{u}$ (eqn S1). I refer to TC as co-dispersion rather than covariance, because traits are treated as fixed in CWM analysis. TC quantifies dispersion of individual traits and co-dispersion between traits in the same manner as would be described by a covariance for a random variable. Just as TN is a baseline for effects of species diversity on trait norms, TC is a baseline for evaluating effects of species diversity on trait dispersion. The abundance-weighted trait dispersion for sample $i$ requires $M$ by $S$ matrices of weights and traits. The weight matrix for observation $i$ is

$$ W_i = \mathbf{1}_m \mathbf{w}'_i $$

The weighted trait co-dispersion is

$$ WTC_i = [(\mathbf{T} - \overline{\mathbf{U}}) \circ W_i](\mathbf{T} - \overline{\mathbf{U}})' $$

where $\circ$ is the Hadamard product. $WTC_i$ is a $M$ by $M$ matrix. Analogous to a coefficient of variation, the square root of the diagonal of the $WTC_i$ matrix can be divided by TN as an index of dispersion for a location (Fig. 1B). Trait co-dispersion over $n$ observations requires the $n$ by $S$ matrix $\mathbf{W}$,

$$ WTC = n^{-1} \left[ (\mathbf{T} - \overline{\mathbf{U}}) \circ \mathbf{1}_m \mathbf{1}'_n \mathbf{W} \right] (\mathbf{T} - \overline{\mathbf{U}})' $$

$WTC$ differs from Laliberté and Legendre’s (2010) functional trait dispersion (FDis) in that $WTC$ is defined on a variance ($\ell^2$) scale rather than a linear scale. I adopt $WTC$ for two reasons. First, it is consistent with the weighted mean for trait norm, with $WTC_i$ being the corresponding weighted norm. Second, it allows direct comparison with $WTC$ that emerges from a model-based approach to species and trait diversity, discussed below. Still, $WTC$ is expected to show patterns similar to FDis. As with the $WTC_i$ this dispersion is induced by the sample $\mathbf{W}$; there is not yet a model.

**Trait prediction for environmental covariance**

The effect of environmental variables is available using a predictive approach. Based on the fitted model in eqn 6 we can predict a realization having the same structure as observed data, starting from the joint distribution for species weights and predictors

$$ [\mathbf{w}_i, \mathbf{x}'_i] = [\mathbf{w}_i | \mathbf{x}'_i][\mathbf{x}'_i] $$

The first factor on the right-hand side is eqn 6. The second factor is a distribution for predictors with mean and covariance,

$$ [\mathbf{x}'_i] = MVN(\mathbf{m}, \mathbf{Q}) $$
Marginalizing $x'_i$ followed by a variable change from $w_i$ to $u_i$ gives a predictive trait distribution

$$u'_i \sim MVN(\text{PM}, \text{PTC})$$  \hfill S6

having predictive mean and predictive trait covariance

$$\text{PM} = T\beta'm$$  \hfill S7

$$\text{PTC} = T(\Sigma + \beta'Q\beta)T'$$  \hfill S8

PTC is the model-based equivalent of WTC in eqn S5. The two terms in eqn S8 are i) the covariance in trait diversity unrelated to environmental predictors, and ii) the covariance that is related to the environment. For the PTM this second term is the environmental trait covariance,

$$\text{ETC} = T\beta'Q\beta'T'$$  \hfill S9a

As a relative contribution of the environmental predictors to the variance we have the ratio

$$\text{RETC} = \frac{\text{diag}(T\beta'Q\beta'T')}{[\text{diag}(\Sigma T') + \text{diag}(T\beta'Q\beta'T')]}$$  \hfill S9b

Because the PTM is fitted to species before translating to traits we can evaluate the same relationships for species, $\text{ESC} = \beta'Q\beta$. The environmental contribution to variance in traits can be compared to its contribution to species

$$\text{RESC} = \frac{\text{diag}(\beta'Q\beta)}{[\text{diag}(\Sigma)+\text{diag}(\beta'Q\beta)]}$$  \hfill S10

Full covariance matrices can be compared using standard tools such as cluster analysis.

Species and trait diversity can be summarized by volume of the hyper-ellipsoid in trait space, taken here at a probability value of 0.05,

$$V(C) = \left[\frac{2(\pi\chi^2_M(0.05))^{M/2} \det(C)^{1/2}}{M\Gamma(M/2)}\right]^{1/M}$$

where $C$ is a $M$-dimensional covariance matrix, $\det$ is the determinant, $\chi^2_M(0.05)$ is the chi-square distribution, and $\Gamma$ is the gamma function. Volumes are multiplicative in dimension $M$; the exponent $1/M$ places this volume on a per-trait basis.
Example Data

The example uses FIA data extracted from 23,916 forested non-plantation plots in 31 eastern states from FIADB version 5.1 on May 5, 2012 (http://fia.fs.fed.us/). All trees with a diameter at breast height (DBH) of at least 12.7 cm were measured on four 7.2 m radius subplots.

The predictors for the FIA data include climate variables, soils, and local drainage. Environmental predictors include a moisture index from the FIA and climate data extracted from the 800 m resolution Parameter-elevation Regressions on Independent Slopes Model (PRISM) data set (http://prism.nacse.org/) (Daly et al. 2009). I used climate data corresponding to each FIA plot to evaluate energy and moisture availability and stress. From mean monthly temperature (°C) and monthly precipitation (mm) I computed the hydrothermal deficit HTD (degree hours with a negative water balance, P < PET), and winter temperature (Clark et al. 2015). To evaluate collinearity I computed variance inflation factors (VIFs) and correlations presented in Table S1.

**Stand age** is recorded as part of the FIA census for each plot.

**Slope and aspect** are captured by three variables,

\[
\begin{bmatrix}
  u_1 \\
  u_2 \\
  u_3
\end{bmatrix} = \begin{bmatrix}
  \sin(slope_i) \\
  \mathbf{sin}(slope_i)\mathbf{sin}(aspect_i) \\
  \mathbf{sin}(slope_i)\mathbf{cos}(aspect_i)
\end{bmatrix}
\]

There is a main effect of slope (first element), but there can be no main effect of aspect independent of slope. If the first coefficient is not different from zero, then slope does not affect the response. If both of the last two coefficients are not different from zero then aspect does not affect the response (Clark 1990).

Table S1. Variance inflation and variable correlation in the design matrix.

<table>
<thead>
<tr>
<th></th>
<th>temp</th>
<th>u1</th>
<th>u2</th>
<th>u3</th>
<th>stand age</th>
<th>moisture</th>
<th>deficit</th>
<th>UltKan</th>
<th>SpodHist</th>
<th>Mol</th>
<th>EntVert</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>1.7</td>
<td>1.3</td>
<td>1.0</td>
<td>1.0</td>
<td>1.10</td>
<td>1.40</td>
<td>1.5</td>
<td>1.2</td>
<td>1.4</td>
<td>1.5</td>
<td>1.1</td>
</tr>
<tr>
<td>factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u1</td>
<td>0.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u2</td>
<td>0.01</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u3</td>
<td>-0.04</td>
<td>-0.13</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>stand age</td>
<td>-0.19</td>
<td>0.23</td>
<td>0.00</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>moisture</td>
<td>-0.03</td>
<td>-0.04</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>deficit</td>
<td>0.49</td>
<td>-0.21</td>
<td>0.01</td>
<td>0.04</td>
<td>-0.24</td>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UltKan</td>
<td>0.19</td>
<td>0.02</td>
<td>0.00</td>
<td>-0.03</td>
<td>-0.10</td>
<td>0.06</td>
<td>0.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SpodHist</td>
<td>-0.48</td>
<td>-0.19</td>
<td>0.00</td>
<td>0.04</td>
<td>0.11</td>
<td>-0.04</td>
<td>-0.30</td>
<td>-0.10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mol</td>
<td>-0.04</td>
<td>-0.03</td>
<td>0.01</td>
<td>0.03</td>
<td>-0.02</td>
<td>0.10</td>
<td>0.13</td>
<td>-0.03</td>
<td>-0.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EntVert</td>
<td>0.06</td>
<td>-0.16</td>
<td>0.00</td>
<td>0.05</td>
<td>-0.06</td>
<td>-0.01</td>
<td>0.13</td>
<td>-0.05</td>
<td>-0.14</td>
<td>0.04</td>
<td></td>
</tr>
</tbody>
</table>
**Soil types** constitute a factor with five levels, Entisol/Vertisol, Mollisol, Spodosol/Histosol, Ultisol-Udults-Kanhapludults (Piedmont Plateau soils), and remaining Ultisols. Data were extracted from the State Soil Geographic (STATSGO) database. For each FIA plot, soil types are categorized based on the first to third soil taxonomy orders. To limit the number of levels in the model similar first order classes were merged into these categories:

Entisol: poor profile development  
Vertisol: plastic clays  
Mollisol: former prairie rich soils  
Spodosol-Histosol: moist, organic, nutrient-poor, low pH,  
Ultisol-Udults-Kanhapludults: clay rich, dominant on the southeastern Piedmont

**Trait data** discussed here include wood density, seed mass, maximum height, leaf [N], leaf [P], specific leaf area, three leaf habit variables, and three ordinal tolerance variables. They are included as a Supplement file. Foliar traits were compiled from Aber and Martin (1999), Alonso et al. (2010), Beaudet, and Messier (2008), Chapin and Kedrowski (1983), Finzi et al. (2001), Gilmore et al. 1995, Grothkopf, and Rejmanek. 2007, LeBauer et al. (2010), Ricklefs, and Matthew (1982), Reich et al. (1998), Rieske (2002), USDA (2015), Van Sambeek et al. (2008), Vergutz et al. (2012), Walker et al. (2014). Leaf habit variables are categorical (0 or 1), including broadleaf deciduous, broadleaf evergreen, and needleleaf evergreen. Drought, shade, and flood tolerances are ordinal data (Niinemets and Valladares 2006, Russell et al. 2014). All others are continuous variables and were centered and standardized for analysis. Seed mass was first transformed to a log scale. Ninety-eight dominant tree species were included with remaining species in FIA gathered in a remainder category. Weights are species abundances, represented as biomass for each sample obtained by allometric conversion of diameter data (Schliep et al., 2015).

**Computation in GJAM**  
Posterior simulation is done by Gibbs sampling, as detailed for generalized joint attribute modeling (GJAM) in Clark et al. (2014, 2015a). GJAM is available for R on CRAN and fully described here. Sampling the GJAM algorithm for the data types included in this model is summarized here.

Regression coefficients for the predictive trait model (PTM) are sampled from

\[
\prod_{i=1}^{n} MVN(\mathbf{w}_i | \mathbf{\beta} \mathbf{x}_i, \Sigma) \ MVN(\text{vec}(\mathbf{\beta}) | 0, \Sigma \otimes \mathbf{C})
\]

where prior covariance \( \mathbf{C} \) is assumed to have infinite variance. The covariance matrix is updated from

\[
\prod_{i=1}^{n} MVN(\mathbf{w}_i | \mathbf{\beta} \mathbf{x}_i, \Sigma) \ |\Sigma|^{-(S+1)/2}
\]
Latent $w_i$ are sampled from the truncated normal distribution subject to a partition that differs for each data type. For ordinal tolerance classes the partition is also sampled. Each is described here. For the trait response model (TRM) substitute $u_i$, $\alpha$, $\Omega$, and $M$ for $w_i$, $\beta$, $\Sigma$, and $S$.

**Additional references**


LeBauer, David; Dietze, Michael; Kooper, Rob; Long, Steven; Mulrooney, Patrick; Rohde, Gareth Scott; Wang, Dan; (2010): Biofuel Ecophysiological Traits and Yields Database (BETYdb); Energy Biosciences Institute, University of Illinois at Urbana-Champaign. http://dx.doi.org/10.13012/J8H41PB9


