This model is wrong. This model is useful.

The goal of this multi-species occupancy model is to simultaneously estimate:

1. Spatial variation in species occupancy along environmental gradients.
2. Spatial variation in site-level species richness along environmental gradients.
3. Spatial variation in pairwise dissimilarity in community composition along environmental gradients.

Spatial variation in species occupancy along environmental gradients

We could just use a multi-species occupancy model for this! To be brief, we go out to i in $1, \cdots, I$ sites to detect s in $1, \cdots, S$ species over j in $1, \cdots, J$ repeat surveys. Let $w_{i,s}$ be the number of surveys we detected each species and $z_{i,s}$ be the latent species presence of species s at site i. We model species presence as a Bernoulli rendom variable such that:

$$z_{i,s} \sim \text{Bernoulli}(\psi_{i,s})$$

where

$$\text{logit}(\psi_{i,s}) = \beta_s^{\psi} x_i$$

Here, we assume there are species level random effects for all intercept and slope terms (e.g., the intercept would look something like $\beta_{s,1}^{\psi} \sim \text{N}(\mu_1^{\psi}, \sigma_1^{\psi})$ where the 1 indicates that this is the intercept, $\mu_1^{\psi} \sim \text{Normal}(0, 100)$ and $\sigma_1^{\psi} \sim \text{Inv-Gamma}(1, 1)$).

The detection model, assuming equal sampling among sites, is:

$$w_{i,s} | z_{i,s} \sim \text{Binomial}(J, \rho_{i,s} \times z_{i,s})$$

Where in this case we can use site-specific covariates on $\rho_{i,s}$ such that:

$$\text{logit}(\rho_{i,s}) = \beta_s^{\rho} x_i$$

And the coefficients can have the same random effect structure. Note that while I'm using x_i in the latent state and detection models, it's only for convenience. They can have different covariates.

Estimating spatial variation in alpha diversity

Often times, alpha and beta diversity can be of more interest, yet there occupancy models do not especially provide a way to evaluate these patterns. There are, of course, standard approaches that ecologists take. To quantify spatial variation in alpha diversity, for example, using some form of regression (e.g., Poisson, Negative Binomial,etc.) could be adequate. Let r_i be the richness of species at site i, which we derive as the row sum of $z_{i,s}$.

We could model detection corrected alpha diversity as:

$$\sum_{s=1}^{S} \boldsymbol{z_i} = r_i \sim \text{Poisson}(\lambda_i)$$

where

$$\log(\lambda_i) = \beta^{\lambda} x_i$$

Giving the coefficients vague priors (e.g., $\beta^{\lambda} \sim \text{Normal}(0, 100)$).

Estimating Beta diversity

Quantifying spatial varition in beta diversity is more complicated, but in general there are two basic approaches. The first is the 'raw-data approach', wherein the enivonmental / geographical variation in beta diversity is partitioned through some form of canonical analysis (e.g., redundancy analysis). The second is a "distance based" approach, where pairwise dissimilarity among sampled locations is measured and then some form of matrix regression is used to correlate these distances to environmental / geographical variation.

As a fan of regression based approaches, I was drawn a towards the distance based approach, as it could theoretically be possible to fold in such a technique into `JAGS`. Of those distance based approaches, generalized dissimilarity modeling (gdm) looked promising. Briefly, after calculating some type of pairwise dissimilarity metric ($d_{i,j}$) for site pair i and j, the regression (using non-negative least squares) looks something like:

$$d_{i,j} = a_0 + \sum_{p=1}^{n} a_p |x_{p,i} - x_{p,j}|$$

where a_0 is an intercept, a_p are slope terms, and $|x_{p,i} - x_{p,j}|$ is the abosolute value of some form of spatial or environmental distance between sites. All coefficients in this model are strictly non-negative. In practice, this model actually uses i-spline basis functions instead of raw covariate values. I-splines range from 0 to 1 and create multiple "new" covariates for each "raw" covariate. These transformed covariates always range between 0 and 1 and steadily creep upwards from 0 to 1 in different, monotonically increasing patterns. Because of this, i-splines ensures that beta diversity increases with environmental distance (i.e., monotonicity), but is still very flexible. In other words, generalized dissimilarity models assume that compositional similarity increases with environmental distance. The `gdm` package in `R` has a lot of the functions to make i-splines and the like (written in C++). I took the time to rewrite them in R while I was working through all of this.

A similar modeling approach to standard gdm, which I thought would be easier to fold into an occupancy model, is to use binomial regression instead. Let $y_{i,g}$ be the number of dissimilar species among a pair of sites and $n_{i,j}$ be the total number of unique species among a pair of sites. Therefore $y_{i,j}/n_{i,j}$ is the proportion of dissimilar species among site pairs (which is essentially Jaccard's dissimilarity). We could then model dissimilarity as:

$$y_{i,g} \sim \text{Binomial}(n_{i,g}, \pi_{i,g})$$

where $\pi i, g$ is the expected dissimilarity.

Then let:

$$\text{logit}(\pi_{i,g}) = a_0 + \sum_{p=1}^{n} a_p (x_{p,g} - x_{p,g})$$

where a_p are given priors on the log scale to ensure they are non-negative (and again we are using i-splines here instead of raw covariates). Conversely, a_0 can be any value on the logit-scale.

Fitting the model in `JAGS`

We could, if we wanted to, fit only the multi-species model and generate some psuedo-posteriors by fitting a frequentist alpha diversity model and beta diversity model to each mcmc step of the $\boldsymbol{Z}$ matrix. There are other papers we could cite as well that have done similar things (though that does not mean it is the correct

thing to do). Instead, I wanted to fit all of these at the same time within JAGS, and condition r_i on $\boldsymbol{z_i}$ and $d_{i,g}$ on $\boldsymbol{z_i}$ and $\boldsymbol{z_g}$. However, to derive r_i and $d_{i,g}$ in JAGS means we need to specify them as deterministic nodes (i.e., <-), but they must also be stochastic nodes (i.e., ~) in JAGS if we want them to estimate spatial variation (i.e., have them be a function of covariates). To address this, I used the Bernoulli one's trick so that the data is input into the hand-coded likelihood (which avoids therefore avoids the need to specify r_i and $d_{i,g}$ as stochastic nodes). I coded up the model in JAGS like so, using nested indexing for $d_{i,g}$ to make sure we evalulate every site pair combination once.

```
model{
  #------------------------------
  # multi-species occupancy model
  #------------------------------
  for(site in 1:nsite){
    for(species in 1:nspecies){
      # Linear predictor latent state model
      logit(psi[site,species]) <- inprod(
        beta_psi[species,],
        design_matrix_psi[site,]
      )
      # latent state z is a Bernoulli random variable
      z[site,species] ~ dbern(
        psi[site,species]
      )
      # Linear predictor data model
      logit(rho[site,species]) <- inprod(
        beta_rho[species,],
        design_matrix_rho[site,]
      )
      # w (observed data) is a binomial process
      w[site,species] ~ dbin(
        rho[site,species] * z[site,species],
        nsurvey
      )
    }
  }
  #----------------------
  # Alpha diversity model
  #----------------------
  for(site in 1:nsite){
    # Derive species richness
    r[site] <- sum(z[site,])
    # linear predictor
    log(mu_alpha[site]) <- inprod(
      beta_alpha,
      design_matrix_alpha[site,]
      )
    # Poisson likelihood
    alpha_lik[site] <- -mu_alpha[site] +
      r[site]*log(mu_alpha[site]) - logfact(r[site])
    # rich is a Poisson random variable via ones trick
    alpha_ones[site] ~ dpois(exp(alpha_lik[site])/CONSTANT)
  }
  #---------------------
  # Beta diversity model
```

```
#---------------------
for(i in 1:n){ # n = number of unique site pair combos
  # Get number of dissimilar species between site pairs (y)
  y1[i] <- sum(
    (1 - z[siteA_id[i],]) *
      z[siteB_id[i],]
  )
  # Get total richness between site pairs (n)
  y2[i] <- nspecies - sum(
    (1 - z[siteA_id[i],]) *
      (1 - z[siteB_id[i],])
  )
  # Linear predictor
  logit(pi[i]) <- b0 + inprod(
    beta_beta,
    design_matrix_beta[i,]
  )
  # code up binomial log likelihood. JAGS does not have a binomial
  # coefficient so it's a bit of a beast to code up. First line
  #  is literally just the binomial coefficient.
  beta_lik[i] <- (logfact(y2[i]) - (logfact(y1[i]) + logfact(y2[i] - y1[i]))) +
    (y1[i] * log(pi[i])) + ((y2[i] - y1[i]) * log(1 - pi[i]))
  # y1 is a binomial process via ones trick
  beta_ones[i] ~ dbern(
    exp(beta_lik[i])/CONSTANT
  )
}
#-------
# priors
#-------
# Multispecies occupancy
for(psii in 1:npar_psi){
  # community mu & sd occupancy
  beta_psi_mu[psii] ~ dlogis(0,1)
  tau_psi[psii] ~ dgamma(0.001,0.001)
  sd_psi[psii] <- 1 / sqrt(tau_psi[psii])
  # community my & sd detection
  beta_rho_mu[psii] ~ dlogis(0,1)
  tau_rho[psii] ~ dgamma(0.001,0.001)
  sd_rho[psii] <- 1 / sqrt(tau_rho[psii])
  # Species specific coefficients
  for(species in 1:nspecies){
    beta_psi[species,psii] ~ dnorm(
      beta_psi_mu[psii],
      tau_psi[psii]
    )
    beta_rho[species,psii] ~ dnorm(
      beta_rho_mu[psii],
      tau_rho[psii]
    )
  }
}
# Alpha diversity
```

```
  for(alphai in 1:npar_alpha){
    beta_alpha[alphai] ~ dnorm(0,0.01)
  }
  # beta diversity
  for(betai in 1:npar_beta){
    beta_log[betai] ~ dnorm(0, 0.01)
    beta_beta[betai] <-  exp(beta_log[betai])
  }
  b0 ~ dlogis(0,1) # beta diversity intercept
}
```

Comparing alpha diversity outputs to other techniques

To do this, I simualted data for 50 sites and 30 species with the multi-species model and a single environmental gradient that the community, on average, responded positively too. As such, we should expect species richness to increase along this environmental gradient (and it does). Having 30 species in this case ensured that there was always at least 1 species present (because dealing with this is a whole other can of worms).

To compare the alpha the outputs from this Bayesian model I fit two other models to the raw data (which is what we would have done if we did not use this model):

1.A "Truth" model. A Poisson glm fit to the simulated (but unknown) true species richness at these sites as a function of the environmental gradient (i.e., `glm(r ~ x, family = "poisson")`). This isn't absolute truth, of course, but it's the best fit line based on the data that is "impossible" for us to observe. 2. An "Observed" model. A Poisson glm fit to the simulated but observed species richness (i.e,. there was imperfect detection). Note that I specified that community level detectability was lower with increasing x. In other words, while species richness went up with x, detectability went down.

And here is what the results look like:

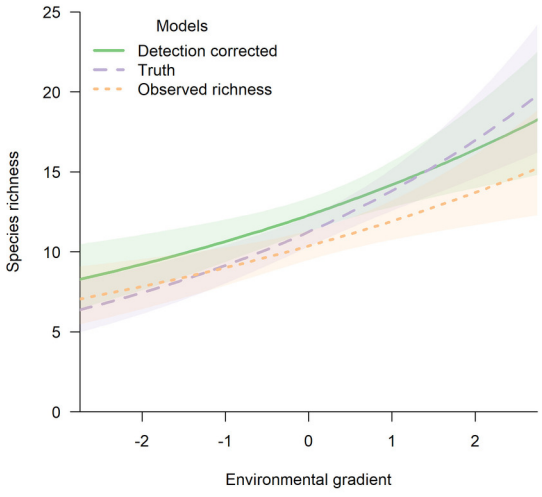

Figure 1: alpha diversity model comparison

Based on this single simulation, the detection corrected model slightly overestimates species richness at the negative end of the environmental gradient. If x = -2.75, the Bayesian model estimated species richness as 8.3 (95% CI = 6.5 - 10.5), the truth model estimated species richness as 6.4 (95% CI = 5.0 - 8.2), and the observed model estimated species richness as 7.1 (95% CI = 5.5, 9.1). Credible interval width was

highest for the detection corrected model (about 4, 3.2, and 3.6 respectively for the detection corrected, truth, and observed models). At the positive end of this gradient (x = 2.75), the detection corrected model slightly underestimated species richness by about 1 species from the true model. Credible interval width was comparable between the detection corrected model and the true model. The observed model performed poorly at positive values of the environmental gradient (probably owing to the fact that detection probability was lower on average with positive values of the gradient).

Comparing beta diversity outputs to other techniques

Comparing beta diversity to frequentist models of a gdm is more difficult because you cannot fit a standard glm (i.e., the likelihood function is different because of the logged slope terms). Fortunately, the (now defunctional) `bbgdm` package has the exact model I wanted to use, so I cannibalized a bunch of the code there to create the frequentist gdm model that uses binomial regression. So again, we have the gdm fit via the bayesian model, a "truth" model, and an "observed" model for comparison. To get confidence intervals and the like I applied a bootstrap (swapping pairs of site in the z matrix). Since there are 50 sites there are only 1225 possible combinations of pairwise site permutations, thus I just did all of them.

A standard way to demonstrate change with gdms is to make a prediction on the link scale (i.e., logit) dissimilarity on the y axis and the environmental gradient on the x axis. This is for one covariate (which was 3 splines in the model), and excludes the model intercept.

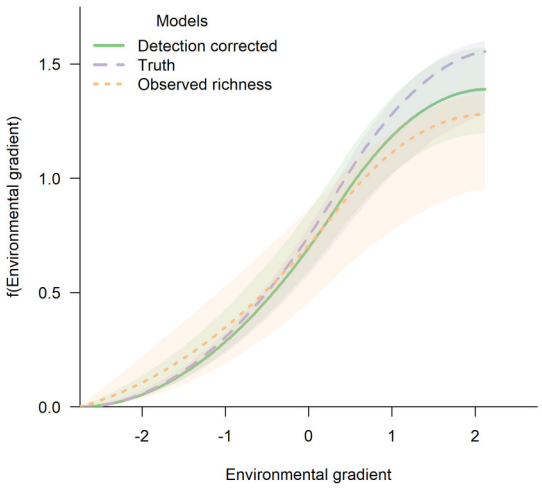

Figure 2: beta diversity model comparison

So what do we see? That results from detection corrected model is more similar to the "true" model than the "observed" model, which is what we'd expect. It seems as if the detection corrected model (in this simulation) slightly underpredicted dissimilarity at the positive end of the gradient, but not by much. 95% CI width were a comperable among models, and the "truth" (i.e,. the beta dissimilarity estimated from the model fit to the true data) fell within the 95% CI of the Bayesian model but not the observed richness model.

My conclusions on this model

Even with the triple dipping we are doing here, this model does better than what we would normally do: fit the models to the observed data and hope for the best. So is this model useful? Yes. Is this model also wrong because it assumes independence among models? Also yes. But, and this is the big but, is that it gets us closer results to "truth" than the standard models that would be fit.

Moving forward

I still wonder if we need to model ψ and ρ within the same linear predictor? This could create the same problem as before, but we could derive expected occupancy as $\psi \times (1 - (1 - \rho)^J)$ (i.e,. the probability of occupancy multiplied by the probability you detect the species at least once in J surveys)? Anyways, now that I have this whole idea flushed out on paper I'd love to hear your thoughts!