Partitioning Abiotic and Biotic Contributions to Community Variation

A Thesis

Submitted to the Faculty

of

Drexel University

by

Steven D. Essinger

in partial fulfillment of the

requirements for the degree

of

Doctor of Philosophy

December 2013

Office of Graduate Studies

Dissertation / Thesis Approval Form

This form is for use by all doctoral and master's students with a dissertation/thesis requirement. Please print clearly as the library will bind a copy of this form with each copy of the dissertation/thesis. All doctoral dissertations must conform to university format requirements, which is the responsibility of the student and supervising professor. Students should obtain a copy of the Thesis Manual located on the library website.

Dissertation/Thesis Title: Partitioning Abiotic and Biotic Contributions to Community Variation

Author: Steven D. Essinger

This dissertation/thesis is hereby accepted and approved.

Signatures:

Examining Committee

Chair

Members

Academic Advisor

Department Head

Office of Graduate Studies • 3141 Chestnut St. • Randell Hall 240 • Philadelphia, PA 19104
Tel.: 215-895-0366 • **Fax**: 215-895-0495 • **Email**: graduate@drexel.edu • **Web**: www.drexel.edu/provost/graduatestudies

Acknowledgments

First and foremost, I would like to thank my advisor, Professor Gail L. Rosen, for her guidance and commitment to my interdisciplinary education over the past five and a half years. She has provided an unexpected education for which I am truly thankful. And to my co-advisor, Professor Christopher B. Blackwood, I am eternally grateful for his interdisciplinary support and willingness to collaborate remotely at Kent State University. He has worn many hats during my thesis research and this dissertation would not have been possible without his dedication. I would also like to extend my thanks and gratitude to the additional members of my Ph.D. committee, Professors Jacob A. Russell, Robi Polikar, John MacLaren Walsh and Leon Hrebien. Their support, guidance and commentary have been instrumental in making this thesis a reality.

To the members of the Ecological and Evolutionary Signal Processing and Informatics Laboratory, Cricket Reichenberger, Greg Ditzler, Yemin Lan, Non Yok, Calvin Morrison and Jean-Luc Bouchot, I offer my sincerest appreciation for your support and camaraderie. And thank you to the "Ecology Crew", whom, in no particular order, Drew, Pat, Karen, Steve, Jake, Maggie, Lori, Jack, Shaya, Laura, Kevin, Abby, Samir, Piotr, Noga, Shauna and a cast of many more, have made an interdisciplinary graduate experience that was as socially stimulating as it was academic for me.

I owe my deepest gratitude to my entire family for their unconditional love and support. Their unwavering belief in me has carried me through the roller coaster of graduate school. My parents, Anne and Steve, have done everything in their power to ensure I have grown up to be an educated and thoughtful individual. My brother and sister, Joey and Kelley, have kept me grounded. And my grandparents, Nanny and Pop, who have given me the love, wisdom, support and guidance I needed to earn a doctorate.

To Jules Winters, whom, in no small part, has been there every step of the way. Her love, support and understanding has provided me with a firm foundation to grow upon. I am eternally grateful for her companionship during these years.

Table of Contents

List of Tables

List of Figures

Abstract

Partitioning Abiotic and Biotic Contributions to Community Variation
Steven D. Essinger
Gail L. Rosen, Ph.D.

It is well known that both environmental factors and species interactions structure ecological communities. To study community composition responses to environmental gradients, ordination and regression techniques are typically employed; however, for studying species interactions, methods primarily rely on analyzing patterns of presence/absence. Each of these types of analyses are carried out independently because there is a lack of unified statistical methods for simultaneous analysis of biotic and abiotic factors influencing community composition. This thesis presents a unified method that enables the removal of environmentally explained variation from species responses so that apparent species interactions are not masked or augmented by the abiotic responses, thus partitioning the abiotic/biotic factors.

To achieve a unified method, first, species responses to environmental gradients are removed via a multivariate regression procedure. Second, the residual responses, void of environmentally explained variation, are tested for species interactions using a null model. Third, communities identified with significant interactions are summarized by the average pairwise covariation among the member species. The method can be used to test hypotheses about species interactions when environmental gradients are present and it may be used to calculate percentages of variation explained due to abiotic, biotic and unexplained factors. Via a sensitivity analysis, I demonstrate that sufficient detection ($\sim$ 95%) and false positive rates ($\sim$ 5%) can be achieved under particular site-species ratios, number of environmental gradients, and covariation-to-noise ratios. My method can guarantee a sufficient average false positive rate ($\sim$ 5%) for communities with > 60 samples, up to 500 species and influenced by up to 4 environmental gradients.

Chapter 1: Introduction

Ecology is the study of interactions among organisms and their environment. Organisms located in close spatial proximity to one another form a community. The structure of the community (e.g. diversity, richness, abundance, etc.) is influenced both by the interactions among the organisms (i.e biotic) and the environmental factors (i.e. abiotic) that characterize the region.

The identification of biotic and abiotic factors structuring ecological communities are cornerstones of ecology, but experimental and statistical methods related to these goals have developed along different paths. Much effort has been given to develop methods of ordination that detect gradients in community composition, which may be driven by directly observed factors or implied latent factors [1, 2]. Such methods may be useful for identifying the strongest abiotic force, but say little about the role of biotic influence and regulation. For example, Redundancy Analysis (RDA) is capable of quantifying the proportion of community variation explained by measured abiotic factors [3]. However, it says nothing about the remaining unexplained variation. RDA has been cited over 1900 times indicating that many studies could benefit from assessing the contribution of biotic factors to the remaining unexplained variation [3, 4, 5].

Ecologists test for biotic interactions using a non-parametric statistical approach known as a null model [6]. Null models condense ecological theory into a prediction about the structure of observed ecological communities and represent a parsimonious approach to determine if observed evidence of biotic interaction is more extreme than expected by chance [7]. They were first developed for binary presence/absence matrices and more recently have been extended to abundance matrices, which may be more effective at testing for segregated, aggregated, and random patterns of abundance [8]. These methods, however, do not incorporate environmental data, and thus cannot decouple the abiotic and biotic factors explicitly. In other words, covariation that is detected among species in the community cannot be partitioned between responses to environmental factors or biological interactions.

A recent method modeling species co-occurrence using multivariate logistic regression combines abiotic and biotic factor analysis into a single method, but is only available for community presence/absence data [9]. Abundance based matrices, however, potentially contain more information on species associations than presence/absence matrices and are thought to be better suited to infer patterns of community structure [8]. Even in cases where a presence/absence approach may be useful, the authors who developed the method do not qualify its performance using a broad sensitivity analysis so it is unclear whether the method performs as expected.

1.0.1 Thesis Contributions

This thesis builds upon previous efforts in abiotic factor and biotic interaction analyses to produce the first hybrid method that removes the influence of environmental factors on community abundance data in tests for biotic interactions. The hybrid method allows community variation attributed to abiotic factors and biotic interactions to be partitioned separately. This is a major advancement in ecology since the method provides a rigorous test to quantify currently unexplained variations in community structure attributed to biotic interactions.

The main contributions of this work are as follows:

1. The first method that is able to analyze biotic interactions in the presence of environmental gradients (i.e. abiotic factors) on community abundance data.

2. The first method that calculates the percentage of variation explained due to abiotic, biotic, and unexplained factors on community abundance data.

3. The first investigation and correction of the prediction bias inherent in Redundancy Analysis.

4. The development of a novel pairwise species interaction model to simulate species interactions.

5. The development of a novel null model and test statistic for biotic interaction analysis.

6. The first method that can detect positive or negative community covariation in the presence of environmental gradients.

7. The first comprehensive study to examine the trade-off between the number of sites, species, environmental factors, noise, signal-to-noise ratio, number of covarying species and types of covariation when using both abiotic and biotic analyses.

8. The first hybrid (abiotic/biotic analysis) method to be released as an open-source R package: http://github.com/EESI/NullSens

9. The first method that can guarantee a sufficient average false positive rate ($\sim 5\%$) for communities with > 60 samples, up to 500 species and influenced by up to 4 environmental gradients.

10. The first study to show that leaf litter microbial communities (from a published dataset) may have more biotic than abiotic variation.

1.1 Outline of Thesis

The following chapters proceed as follows:

- Chapter 2 focuses on the response of organisms to environmental factors. A survey of popular methods for characterizing collective community responses of organisms to environmental factors is provided. A method for simulating community responses to these factors is described. Bias among residual and predicted responses is explored. Alternative approaches to mitigate this bias are presented. An experiment showing the bias among predicted and residual responses for these alternative methods is provided along with identification of the best performing methods.

- Chapter 3 focuses on species interactions within communities. Methods featuring null model approaches for detecting significant interactions among species are presented. A method for quantifying species interactions using a pairwise interaction model is developed along with a method for simulating species interactions.

- Chapter 4 presents sensitivity analyses of the performance of the methods combined from chapters 2 and 3. The analyses are conducted using Monte Carlo simulations of species responses and interactions as described in the previous chapters. Type I errors are summarized

for the collection of methods. The best performing methods are scrutinized further along with characterization of the detection rate of the best performing method.

- Chapter 5 presents a hybrid method combining environmental and biotic interaction analyses. The method is chosen based on performance in the sensitivity analyses of chapter 4. The hybrid method is summarized and a benchmark of the computational complexity is provided.
- Chapter 6 illustrates the application of the hybrid method to a published leaf litter dataset. This case study shows the use of the method on a typical application encountered in practice. The method is shown to quantify 46% of the previously unexplained variation.
- Chapter 7 provides overall conclusions and directions for future research.

Chapter 2: Environmental Responses

2.1 Introduction

This chapter will focus on the response of organisms attributed to environmental (i.e. abiotic) factors. Abiotic factors vary over space and time thus providing for dynamic community structures. Due to the influence on community structure the study of the responses of organisms to perturbations of abiotic factors is a primary interest in ecology [10].

This chapter begins with a survey of popular techniques dedicated to characterizing the collective community responses of organisms to abiotic factors. Particular attention is paid to methods that utilize measured factors over those that infer latent factors since the latter do not facilitate separation of abiotic and biotic influences. Once abiotic responses are characterized the remaining unexplained residual responses are analyzed for suitability for subsequent biotic interaction analyses described in chapter 3. Inherent bias is discovered in the residual responses thus prompting the exploration of alternative approaches. An analysis of potential solutions is explored.

2.2 Background

Methods employed for assessing environmental and community relationships typically search for correlations among community composition and abiotic factors using ordination techniques [11]. Ordination is the spatial representation of objects such that the proximity of objects to one another is based on similarity [12]. Dependent upon the method the axes represent explicit or implicit abiotic and biotic factors. Typically, the objects represent sample sites with each site representing a community and given set of abiotic factors. The similarity is a measure computed among the sites based on the organisms present at each site. It could be as simple as the euclidean distance or of greater complexity such as the non-metric Bray-Curtis statistic [13]. Methods of ordination are characterized as indirect if specific abiotic factors are excluded in the computation and direct if explicitly specified.

2.2.1 Indirect Methods

Indirect methods seek to explain variation among communities without constraint to specific abiotic factors and are sometimes referred to as unconstrained methods [12]. Ordinations may be obtained by finding the dimensions having the greatest variance and projecting the data down to a 2 or 3 dimensional space for visualization. The most basic approach is principal components analysis (PCA) [3]. Extensions of PCA that permit more biologically relevant similarity measures are principal coordinates analysis (PCoA) and non-metric multidimensional scaling (NMDS) [14, 15]. All of the indirect methods describe an ordination with axes referring to some unknown latent factors. Notably these factors indiscriminately summarize both abiotic and biotic contributions. Hypotheses about the relationship of specific factors and the community may be tested via correlation with the ordination, but it is unrealistic to expect partitioning of abiotic and biotic contributions. A survey of indirect methods comparing the performance among multiple similarity measures has been investigated in depth [2].

2.2.2 Direct Methods

Direct methods seek to explain variation among communities by constraining the ordination to combinations of the measured abiotic factors [12]. The axes returned by these methods directly correspond to variation of the explanatory variables. An explicit assessment of the relationship between community structure and the abiotic factors is therefore readily available. Regression based ordination methods facilitate abiotic variance partitioning by assessing the congruence between the observed and predicted responses via the coefficient of determination [16]. Constrained methods have been employed for this purpose, but approaches have stopped short of investigating biotic interactions [17]. Chapter 3 will focus on the use of the regression residuals for testing hypothesis of biotic interactions.

Redundancy Analysis (RDA)

RDA is the preferred method for constrained analysis of species responding linearly to abiotic factors [11]. It is the combination of multivariate multiple linear regression with principal components

analysis (PCA) [3, 18]. Responses are regressed on a set of abiotic explanatory variables via ordinary least squares (OLS) to estimate a set of regression coefficients, from which fitted responses are obtained. A covariance matrix is constructed from the fitted values and is subjected to PCA for the purposes of dimensionality reduction. The ordination of objects (sample sites) can be displayed in a reduced space of the explanatory variables via multiplication of the fitted values with the eigenvectors obtained from the PCA.

Canonical Correspondence Analysis (CCA)

CCA is a technique that approximates a constrained Gaussian ordination [19]. It is appropriate for use on species response data that resembles a unimodal curve such as those encountered with responses measured over a large range of abiotic factor(s) [1, 20]. Mean, variance and height parameters are estimated for each species response along an axis of scores obtained by multivariate regression on abiotic explanatory variables. Among other restrictions the method assumes each species shares the same tolerance (i.e. variance) and preferred condition (i.e. mean) for the measured abiotic factors. While this is unrealistic in practice the method been shown to be somewhat robust to deviations in the restrictions. These deviations, however, lead to an artifact known as the arch effect that obscures the ordination with non-linearities [21].

2.2.3 Community Data

There are a few types of data available for quantifying communities. These include presence-absence (e.g. binary), relative abundance (i.e. compositional) and absolute abundance. The choice and availability of each type is largely dependent upon the particular study and the sampling technology utilized.

Presence-Absence

The most basic form of community data is the presence-absence scheme. The data produces a matrix populated with zeros and ones where a one indicates a particular organism was found at a particular site. This binary data type has been used extensively in species co-occurance studies and

the performance of null model methods on this type of data has been studied extensively [6]. Null models will be discussed in chapter 3.

Relative Abundance

Relative abundances arise from compositional data where the total species abundance sums to one at each sample site. This is the most frequently encountered data type in studies involving microorganisms and is attributed to the sampling detection process. DNA from a community of organisms is extracted and sequenced using technology such as Pyrosequencing [22]. This sampling process allows fosr discrimination among the organisms present in the sample, but does not provide counts on the absolute abundance of each organism. Additional information such as biomass may be collected at each sample site and used to convert the relative abundance to absolute, but this additional data is generally not collected during studies.

Absolute Abundance

Ecological studies involving macroorganisms such as vegetation may rely on direct counting for data collection. This data type provides greater sensitivity in the analysis of communities over the presence-absence scheme since we move from a binary to a continuous domain. Correspondingly, absolute abundances have more recently been investigated for the use of null model approaches in studying species co-occurence [8]. Methods for investigating covariation among species would benefit most from absolute abundance data since relative abundance inherently produces correlations among species responses and may thus provide misleading results.

Truncated Data

Community data contains only positive values and zeros. In practice, datasets typically have many zeros. The notion of negative abundance does not have a clear physical meaning. Negative abundances are not encountered since species counts will never yield negative values. Community data is therefore said to be truncated at zero and presents a challenge in modeling species responses. Truncated data can bias parameter estimation and can lead to artificial linear responses in residuals.

Typically, the bias is ignored in methods such as RDA, but will be addressed here since the analysis of residual responses is important for subsequent analyses.

Interpretations of Zeros

Difficulty lies in determining the origin of zeros in a dataset since there may be more than one plausible cause for each observation. Aside from zeros arising due to sampling processes (i.e. missed detection) the predominant origin of zeros include an unsuitable environmental condition for a species or compensatory dynamics [23]. Compensatory dynamics occur when one species drives another to zero abundance at a given sample site. One or more of these three possibilities may be responsible for the observation and can affect the outcome of abiotic and biotic analyses. Species response models should include zeros to reflect real data and it is desirable for methods to judiciously handle the cause of each zero although currently seldom achieved in practice.

2.3 Modeling Abiotic Responses

The multivariate linear model employed by RDA is an excellent place to start in the development of an abiotic response model due to its vetted acceptance in the research community and simplicity of model parameter estimates [1, 3, 18, 5]. The model will need to be modified, however, as I will show that spurious correlations are introduced among residuals and predicted responses when used with species count data. This issue stemming from bias in RDA has not been investigated prior to this thesis and will be discussed in the subsequent section on analysis of residual responses.

2.3.1 Simulating Observed Responses

The abundance of species, k at a particular site, j is described by equation (2.1). Each species, k responds to a set of environmental factors, $\mathbf{x}_j$ at each site (equation (2.2)). There are a total of n sites, p species and q environmental gradients. Parameters, b_{ki} account for the contribution of each

factor, i to species, k abundance. Values for these variables are sampled from the distributions in equation (2.3).

$$y_{jk} = b_{k0}x_{j0} + b_{k1}x_{j1} + b_{k2}x_{j2} + \cdots + b_{kq}x_{jq}, \quad \begin{cases} j = 1, 2, \ldots, n \; sites \\ k = 1, 2, \ldots, p \; species \\ i = 0, 1, \ldots, q \; gradients \end{cases} \tag{2.1}$$

$$\mathbf{x}_j = [x_{j0}, x_{j1}, x_{j2}, \ldots, x_{jq}] \tag{2.2}$$

$$\begin{gathered} b_{ki} \sim N(0, 1) \\ x_{ji} \sim U(-1, 1), \quad x_{j0} = 1, \; \forall \; j \end{gathered} \tag{2.3}$$

The environmental factor matrix, $\mathbf{X}$ in equation (2.4) can be constructed by including the environmental variables from each sample site in equation (2.2). Similarly, the species parameters, b_{ki} may be vectorized across all factors as shown in equation (2.5). A single species, k responding to all environmental factors over all sites is described by equation (2.6).

$$\mathbf{X} = \begin{bmatrix} \mathbf{x_0} \; \mathbf{x_1} \; \mathbf{x_2} \; \ldots \; \mathbf{x_q} \end{bmatrix} \tag{2.4}$$

$$\boldsymbol{\beta}_k = [b_{k0}, b_{k1}, b_{k2}, \ldots, b_{kq}] \tag{2.5}$$

$$\mathbf{y}_k = \mathbf{X}\boldsymbol{\beta}_k \tag{2.6}$$

A community of species, $\mathbf{Y}$ is formed by combining the respective species parameters, $\boldsymbol{\beta}_k$ as in equation (2.7) and multiplying by the environmental matrix in equation (2.4) as shown in equation (2.8).

$$\mathbf{B} = \begin{bmatrix} \boldsymbol{\beta}_1 \, \boldsymbol{\beta}_2 \, \cdots \, \boldsymbol{\beta}_p \end{bmatrix} \tag{2.7}$$

$$\mathbf{Y} = \mathbf{XB} \tag{2.8}$$

Noise may be added to the species responses to simulate unaccounted variation (equation (2.9)), where the noise is drawn from a normal distribution with variance σ_N^2 (equation (2.10)).

$$\mathbf{Y} = \mathbf{XB} + \mathbf{N} \tag{2.9}$$

$$N_{jk} \sim N(0, \sigma_N^2) \tag{2.10}$$

All values of $\mathbf{Y}$ that are less than zero are zero truncated to reflect species count data as observed in practice (i.e. no negative abundances).

2.3.2 Predicted & Residuals Responses

Species response parameters, $\hat{\boldsymbol{\beta}}_k$ can be estimated from community data (simulated or real) and a set of environmental factors. These parameters are estimated using the normal equation for ordinary least squares (OLS) shown in equation (2.11) or equation (2.12) (matrix form). OLS is the maximum likelihood method for regression parameter estimation [24].

$$\hat{\boldsymbol{\beta}}_k = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}_k \tag{2.11}$$

$$\hat{\boldsymbol{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{2.12}$$

The predicted response, $\hat{\mathbf{y}}$ is the expected species response to the provided environmental factors given the linear model assumption. Predicted (fitted) responses are obtained by multiplying the estimated parameters, $\hat{\boldsymbol{\beta}}_k$ by the explanatory variables, X as shown in equation (2.13) or 2.14 (matrix form).

$$\hat{\mathbf{y}}_k = \mathbf{X}\hat{\boldsymbol{\beta}}_k \tag{2.13}$$

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{B}} \tag{2.14}$$

Residuals are subsequently computed via subtraction of the observed, $\mathbf{y}_k$ from the predicted responses $\hat{\mathbf{y}}_k$ as shown in equation (2.15) or 2.16 (matrix form). The residuals describe observed responses that deviate from those predicted by the model. The residuals therefore should contain all variation unaccounted for by the model and should be void of linear relationships with the environmental factors.

$$\mathbf{y}_{\mathbf{res}k} = \mathbf{y}_k - \hat{\mathbf{y}}_k \tag{2.15}$$

$$\mathbf{Y}_{\mathbf{res}} = \mathbf{Y} - \hat{\mathbf{Y}} \tag{2.16}$$

2.3.3 Quantifying Abiotic Responses

The amount of variation in the species response explained by the environmental factors is typically computed using the coefficient of determination, R^2. This is a measure of the amount of correlation among the observed and predicted responses [16]. There are several methods for computing this value, but it is generally a weighted average based on the abundance of each species within the community.

Standard Approach

Community R^2 is first computed for each species response, k and then averaged over all of the responses. The R^2, 2.20 for each species may be computed using the total sum of squares 2.18 and the residual sum of squares 2.17, where $\hat{y}_j$ is the predicted value at site j and $\bar{y}_k$ is the average abundance for species k over all sites, equation (2.19).

$$SS_{res_k} = \sum_{j=1}^{n} (y_{jk} - \bar{y}_k)^2 \tag{2.17}$$

$$SS_{tot_k} = \sum_{j=1}^{n} (y_{jk} - \hat{y}_{jk})^2 \tag{2.18}$$

$$\bar{y}_k = \frac{1}{n} \sum_{j=1}^{n} y_{jk} \tag{2.19}$$

$$R_k^2 = 1 - \frac{SS_{res_k}}{SS_{tot_k}} \tag{2.20}$$

Weighted Approach

Certain species may be more abundant in the community than others and so their contribution to the community R^2 should be greater. In this case a weighted R^2 approach is favored where the total community R^2 is a weighted average based on the abundance of each species. RDA computes this number by 2.21, where tr() is the trace of the matrix [11]. This is equivalent to the summed ratio of predicted to observed eigenvalues.

$$\frac{tr(Y_{hat} Y_{hat}^T)}{tr(Y Y^T)} \tag{2.21}$$

Alternatively, a weighted R^2 may be computed by equation (2.22), where w_k is the total abundance of the k-th species (equation (2.23)) and W is the sum of abundances over all p species, equation (2.24).

$$R^2_{avg} = \frac{1}{W}\sum_{k=1}^{p} w_k R^2_k \tag{2.22}$$

$$w_k = \sum_{j=1}^{n} y_{jk} \tag{2.23}$$

$$W = \sum_{k=1}^{p} w_k \tag{2.24}$$

Adjusted $\mathbf{R}^2$

Typically, the value of R^2 will increase spuriously with the increasing number of environmental variables considered in the model. Regardless of the type of R^2 coefficient chosen, it is appropriate to compute the adjusted coefficient as shown in 2.25 as it has been shown to be a much more accurate estimate of the true model fit for species count data [17].

$$R^2_{adj} = R^2_{avg} - (1 - R^2_{avg})\frac{q}{n - q - 1} \tag{2.25}$$

2.4 Analysis of Residual Responses

The residual responses remaining after performing multivariate regression should be void of linearity associated with the environmental factors. Testing for significant covariation among the residual species responses is desired to explore potential biotic interactions within the community. It is important that the regression approach does not inadvertently alter the residuals by introducing spurious artifacts that could lead to problems such as false detections (e.g. high Type I errors). This is typically not an issue with standard multivariate regression. However, I have found that the effects of regression on truncated species count data does indeed introduce artifacts leading to spurious correlations among species response residuals and their respective predicted responses. This

section explores the issue further and offers potential solutions to mitigate the potential for high type I error rates due to these artifacts.

2.4.1 Spurious Correlations among Predicted & Residual Responses

Investigation of the effects of regression on the residuals is well suited for study with simulated species response data. This allows control for factors such as noise, explanatory variables and sites in understanding the root cause of the spurious correlations. The simulated species responses used herein are constructed using the approach described previously in this chapter.

A visualization of a species responding linearly to a single environmental factor is shown in figure 2.1. The true linear response is shown in cyan and is described by equation (2.6). Noise has been added to the response to simulate uncertainty in the measurement and the model, thus giving the observed response (blue). The predicted response (green, equation (2.13)) is obtained by estimating the response parameters using equation (2.11) on the observed data (blue). The residual response (red, equation (2.15)) shows that wherever there are simultaneously zeros in the observed response and negative predicted values an inverted predicted response will appear in the residuals. In this particular example the correlation between the predicted and the residual response is one for sites where the predicted response is negative. This is an overall trend that is not exclusive to this example.

Truncating the predicted response at zero for values predicted to be negative helps reduce the correlation among the predicted and residual responses (equation (2.26)). Subsequent analysis shows that the correlation is still problematic, however, even with this additional step. Further inspection of figure 2.2 indicates that the predicted response well over-estimates the true response. Figures 2.3 and 2.4 illustrate the correlation issue in two dimensions. The problem remains persistent in higher dimensional spaces.

The next section shows that negative predicted values should be truncated at zero (equation (2.26)) to reduce correlations among predicted and residual responses. This is acceptable since negative values are not realizable and their inclusion would put artificial linear responses in the

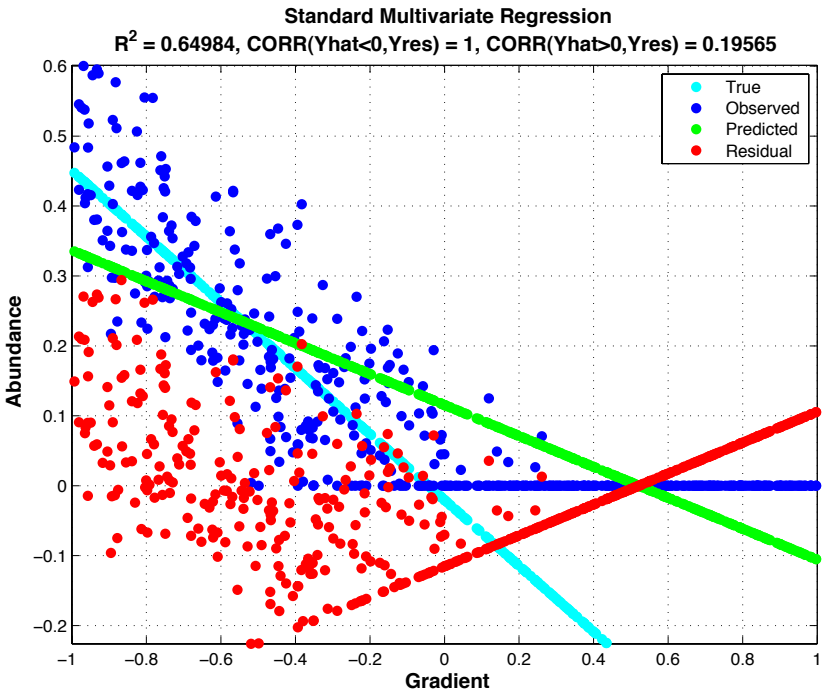

Figure 2.1: Standard multivariate regression on simulated species data. The true, observed, predicted and residual responses are shown.

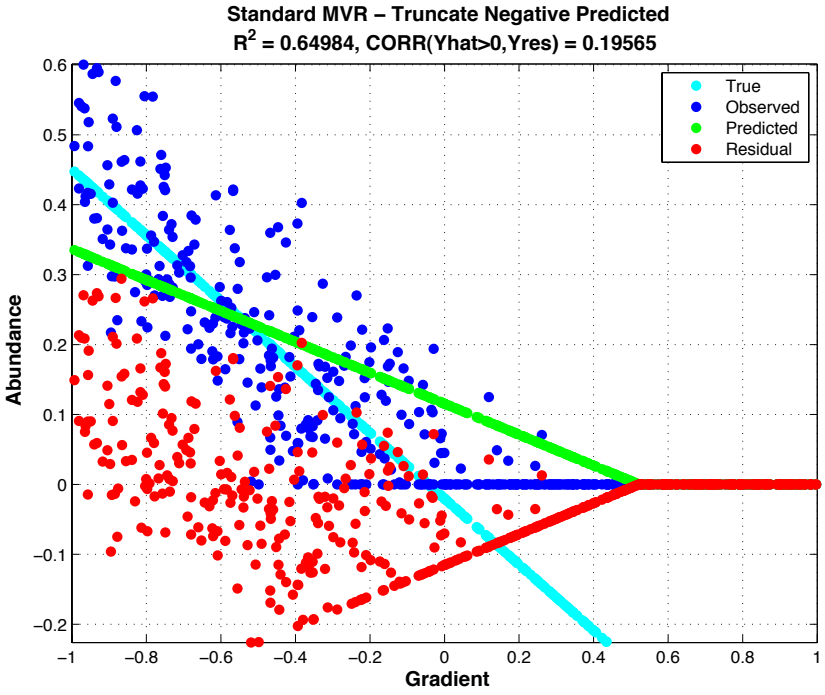

Figure 2.2: Standard multivariate regression on simulated species data. The true, observed, predicted and residual responses are shown. Note that the negative predicted responses are truncated to zero.

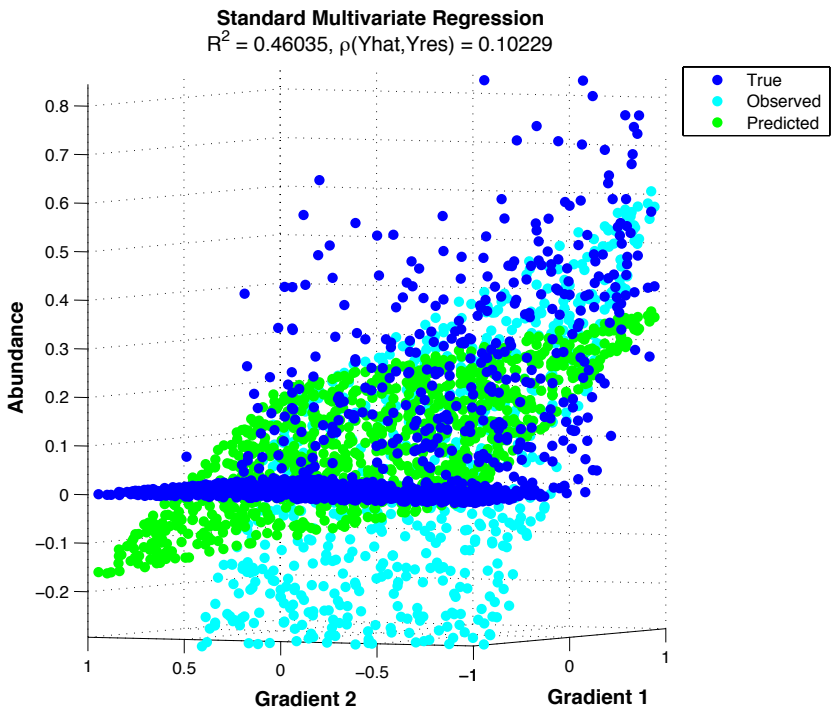

Figure 2.3: Simulated species response to two environmental gradients. True, observed and predicted responses are shown in the figure.

residuals. Note that the truncation of negative abundances is an additional step included here that is not part of traditional RDA.

$$\hat{\mathbf{y}}_k(< 0) = 0 \tag{2.26}$$

These examples illustrate how truncated data such as species counts can bias parameter estimation. Avoiding bias is a challenging endeavor because there are many contributing conditions. For example, in high R^2 (low noise) environments responses smaller in magnitude (e.g. small slope, small effect) tend to be overestimated. Larger responses tend to be underestimated. The result is that linear responses show up in the residuals due to linear deviation of the estimated from the observed response. On the contrary, in low R^2 environments (high noise) the slope of the predicted response approaches zero thus passing a negatively shifted image of the observed responses to the residuals,

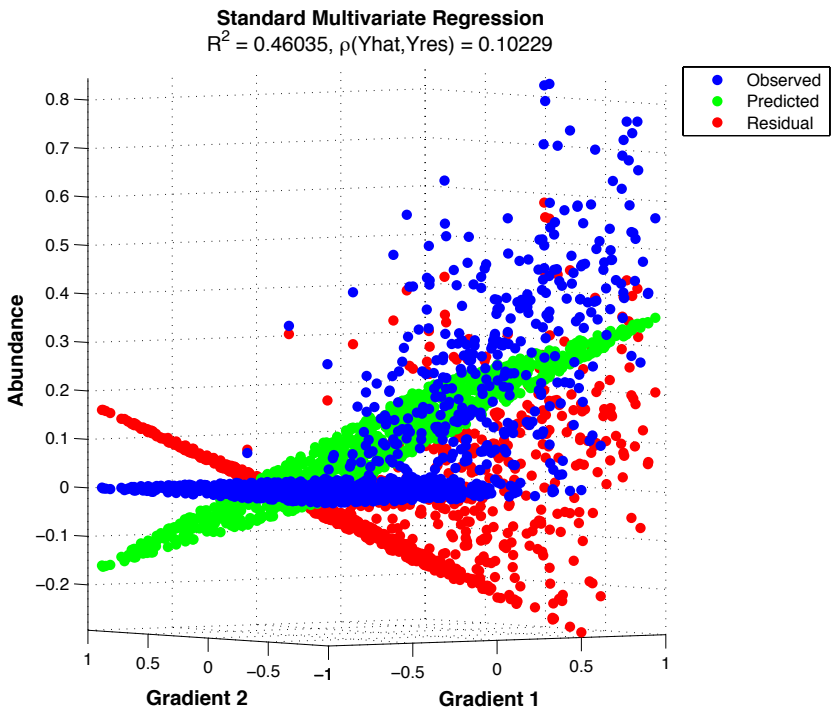

Figure 2.4: Simulated Observed, Predicted and Residual Responses to two Environmental Gradients. Predicted response obtained using standard multivariate Regression.

cutoff by the reflection of the slope of the predicted response. This also introduces correlation in the residuals and is problematic for Type I error rates.

2.4.2 Alternative MVR Approaches

The issues encountered with inverse predicted responses appearing in the residuals prompts the exploration of alternative approaches with the aim of mitigating this undesirable artifact. A few plausible approaches are investigated for replacing the standard multivariate regression to circumvent issues with species count data. The first two approaches deal with robust statistical approaches while the remaining deal with novel data driven methods.

Tobit MVR

Tobit multivariate regression is the maximum likelihood solution for working with censored data [25]. Tobit regression is a class of methods that originated out of the econometrics literature in the 1950's for fitting observations that are clipped (censored) at a minimum or maximum value.

For example, income versus expenditure produces censored data since a positive income may be observed, but the expenditure on a commodity could be zero. Furthermore, a negative expenditure cannot be observed so the data must be censored at zero.

The Tobit framework may be described by equations (2.27) and (2.28). Species abundance observation $\mathbf{y}_i$ can modeled by a latent variable $\mathbf{y}_i^*$ as in (2.27). This latent variable may then be described by a linear model as in standard regression. The parameters can then be estimated using expectation-maximization [26].

$$\mathbf{y}_i = \begin{cases} \mathbf{y}_i^* & \text{if } \mathbf{y}_i^* > 0 \\ 0 & \text{if } \mathbf{y}_i^* \leq 0 \end{cases} \tag{2.27}$$

$$\mathbf{y}_i^* = \boldsymbol{\beta}\mathbf{x}_i + u_i, \quad u_i \sim N(0, \sigma^2) \tag{2.28}$$

Note that the ecological literature usually refers to censored data as truncated data so I will continue to use the latter reference here to remain consistent with the ecology literature.

Robust MVR

Robust multivariate regression has been developed to circumvent issues commonly encountered that violate the assumptions of standard multivariate regression [27]. The least squares loss function employed by standard regression as described by equation (2.29) penalizes all data points based on the squared distance of the point to the fitted line. Therefore, a single outlier can cause the fitted line to favor the outlier since the squared distance is large. Alternatively, the bisquare loss function, equations (2.30) and (2.31) are one of many alternatives to the least square loss function. Figure 2.6 illustrates the difference between these two loss functions. The penalty for data points observed past a threshold is held constant so that these 'outliers' do not pull the fitted line disproportionately

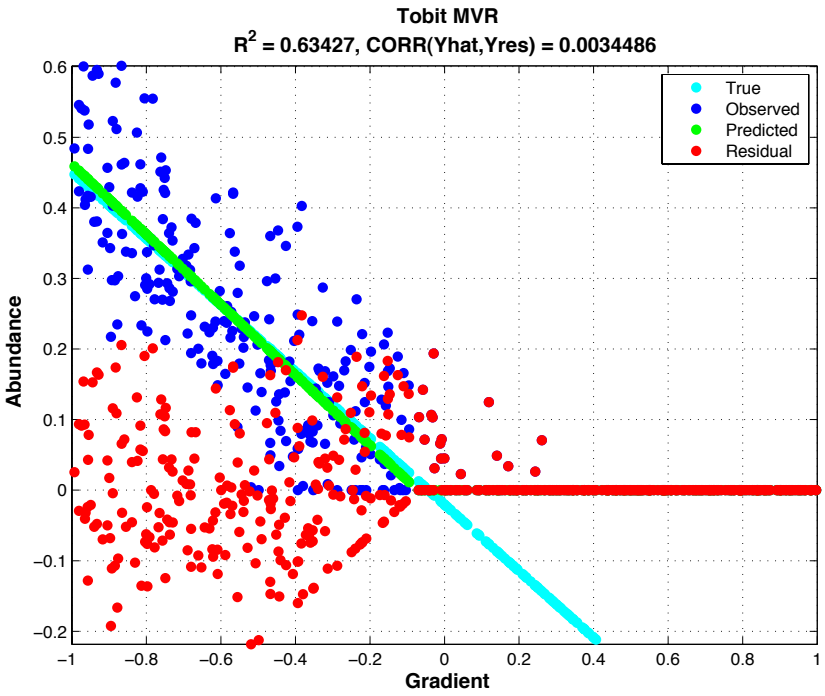

Figure 2.5: Tobit regression on simulated species response data. The true, observed, predicted and residual responses are shown. The predicted response is a much better fit of the true response than obtained with standard multivariate regression. The correlation between the predicted response and the residual is orders of magnitude smaller than with standard multivariate regression on truncated data.

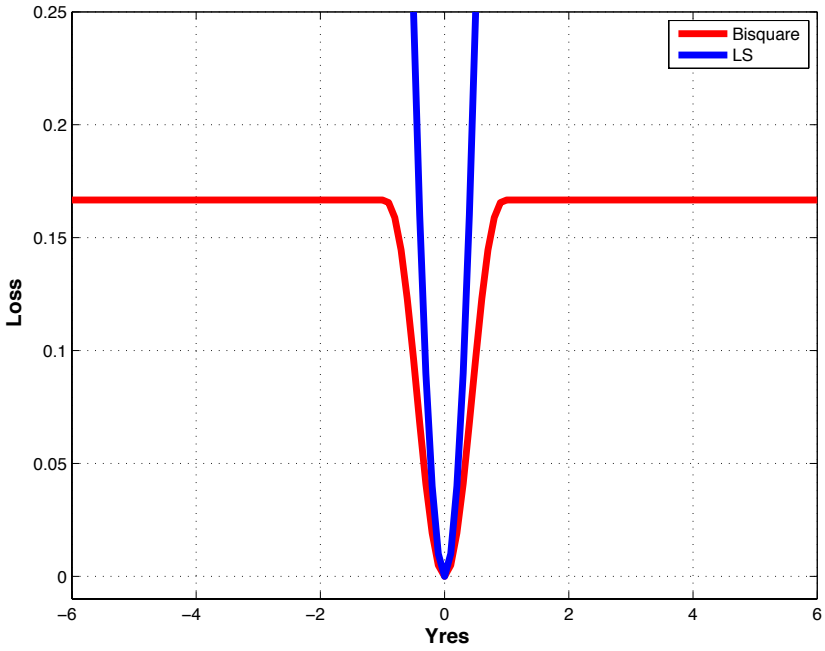

Figure 2.6: Bisquare versus Least-Squares Loss Function. Least-Squares penalizes data by the squared distance from the fitted line to the observation. Alternatively, the bisquare loss function penalizes all points beyond a threshold identically so that outliers do not pull the fitted response away from the majority of observations.

toward the outliers. The result is a potential better fit of the model to the majority of the observed data.

$$L(\mathbf{Y}_{res}) = \sum_{i=1}^{n} \mathbf{Y}_{res_i}^2 \tag{2.29}$$

$$L(\mathbf{Y}_{res}) = \sum_{i=1}^{n} \rho(\mathbf{Y}_{res_i}) \tag{2.30}$$

$$\rho(e) = \begin{cases} \frac{k^2}{6}\left\{1 - \left[1 - (\frac{e}{k})^2\right]^3\right\} & \text{for } |e| \leq k \\ k^2/6 & \text{for } |e| > k \end{cases} \tag{2.31}$$

Site Selective MVR

A site selection procedure employed prior to performing multivariate regression is investigated. The inclusion of all sample sites in regression parameter estimation becomes problematic for correlation analysis of response residuals when working with truncated data such as species counts. Predicted responses derived from these parameter estimates are effectively biased stemming from uncertainty about the origin of zeros observed within the data (e.g. Is the zero due to a missed sampling opportunity or unsupported environmental condition?). Regression bias passes through to the residuals via subtraction of observed from predicted responses (i.e. $\mathbf{y}_k$- $\hat{\mathbf{y}}_k$). This is particularly noticeable when there is simultaneously an observed zero with a negative predicted response at a particular site.

Bias in the residuals leads to spurious correlations among species residual responses thereby potentially increasing the false positive rate of detection to unsatisfactory levels. To mitigate the bias arising from the use of truncated data, the origin of zeros in the response data is herein considered when selecting sites appropriate for inclusion in the regression. The response of each species should be analyzed separately across all sites with each species having a respective set of sites for regression.

Bounding by Positive Abundance

Zeros arising from unsupported environmental conditions are particularly problematic for the false positive rate since non-zero predicted responses at these sites pass directly through to the residuals. This is a likely situation since predicted responses are generally inflated due to estimation on truncated data. Removing sites with zeros falling outside the most extreme positive observed abundances ensures that the sites likely attributed to unsupported environmental conditions for a given species are not included in further analysis. Therefore, a potential remedy for bias includes discarding sites further than the furthest observed non-zero abundance.

The method works by finding all of the observed positive abundances for a respective species as in equation (2.32). The maximum and minimum values for each environmental factor are found using only sites observed with positive abundance as in equation (2.33). Only sites (both positive and zero abundance) within this range are retained as described by equation (2.34). The final

set of sites retained for regression are those in the intersection of the sets of sites retained for each environmental factor as in equation (2.35). This procedure is repeated independently for each species in the community.

$$\boldsymbol{\zeta}_k = \{j \in \{1, 2, \ldots, n\} : \mathbf{y}_k(j) > 0\} \tag{2.32}$$

$$T_{min_i} = \min(\mathbf{x}_i(\boldsymbol{\zeta}_k)), \quad T_{max_i} = \max(\mathbf{x}_i(\boldsymbol{\zeta}_k)) \tag{2.33}$$

$$\mathbf{s}_i = \{j \in \{1, 2, \ldots, n\} : T_{min_i} < \mathbf{x}_i(j) < T_{max_i}\} \tag{2.34}$$

$$\mathbf{S}_k = \bigcap_{i=1}^{q} \mathbf{s}_i, \qquad \mathbf{S}_k \subseteq \{1, 2, ..., n\} \tag{2.35}$$

Figure 2.7 shows an example of sites with positive abundance (black) and zeros (green) in two dimensional space (i.e. two environmental factors). In assuming that species do not exist outside of their environment the zero sites beyond any observed positive abundance may be dropped from the regression. This will reduce the opportunity for the predicted response to show up in the residuals. Figure 2.8 illustrates the new site selected regression region. The zero sites within the region may not be removed from the analysis, however, since doing so would inflate the predicted response and make the detection of negative covariation problematic. The latter is a likely scenario when competitive interactions be take place in the community.

Bounding by Abundance Threshold

As shown in figure 2.8, many zeros can remain between the furthest positive abundance and the centroid of abundances. Therefore, a more aggressive threshold approach may be required since the furthest positive abundance may be an outlier. This outlier may need to be defined more conservatively than in the usual three sigma sense [28]. The bounding by abundance threshold method computes a simple threshold for selecting sites for inclusion. This threshold is the new

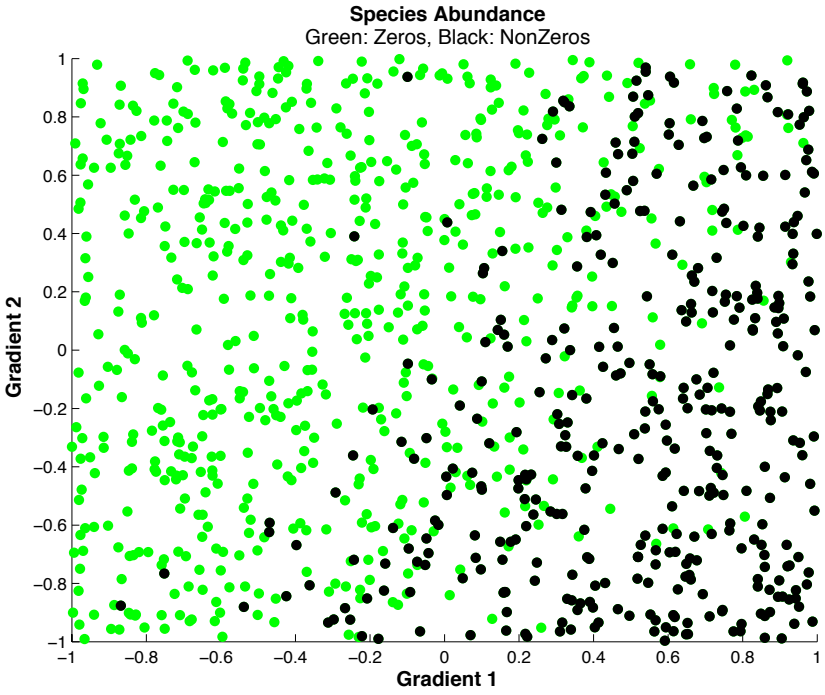

Figure 2.7: Location of zero and positive abundances.

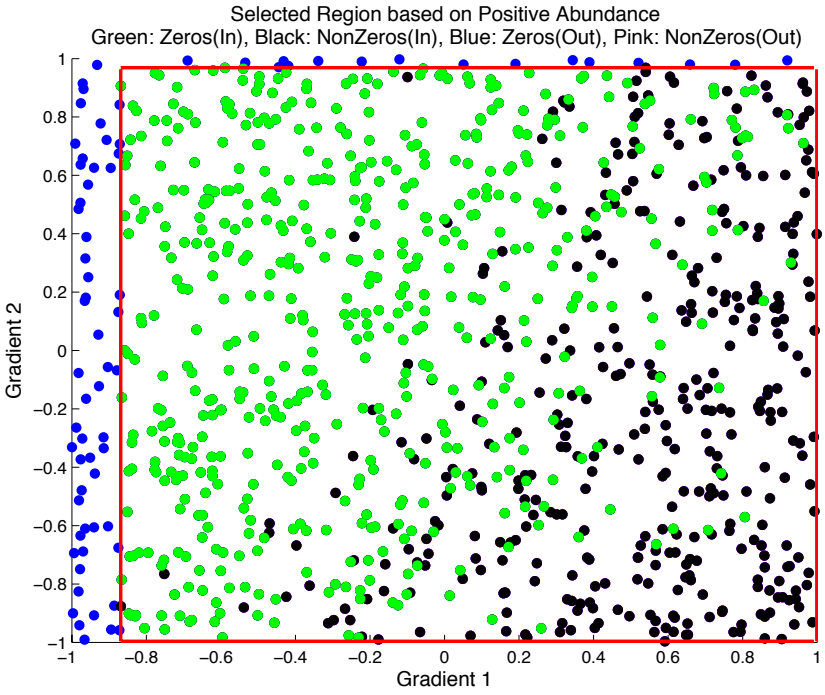

Figure 2.8: Selected regression region based on positive abundances.

definition for an outlier in the dataset and is computed separately for each species. Note that outliers in the three sigma sense should be removed prior to determining the bounding region.

For each explanatory variable the centroid (median) of all positive abundances for the given species is computed, equation (2.37). The average distance of each positive abundance from the centroid is then obtained. A threshold determined by the average distance of positive abundance from the centroid plus one standard deviation is imposed as shown in equation (2.38). All sites beyond the threshold are removed from the regression analysis, while all within (both positive and zeros) are retained as in equation (2.40). The thresholding processes continues until all explanatory variables have been considered each in turn. The final set of sites selected for inclusion is the intersection of sites retained after each thresholding as in equation (2.41). This approach is similar to drawing a hypercube around the sites in environmental variable space and selecting those points inside for inclusion.

$$\boldsymbol{\zeta}_k = \{j \in \{1, 2, \ldots, n\} : \mathbf{y}_k(j) > 0\} \tag{2.36}$$

$$C_i = \tilde{\mathbf{x}}_i(\boldsymbol{\zeta}_k) \tag{2.37}$$

$$T_i = \frac{1}{n}\sum_{j=1}^{n}(x_{ji} - C_i) + \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)} \tag{2.38}$$

$$\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ji} \tag{2.39}$$

$$\mathbf{s}_i = \{j \in \{1, 2, \ldots, n\} : (C_i - T_i) < \mathbf{x}_i(j) < (C_i + T_i)\} \tag{2.40}$$

$$\mathbf{S}_k = \bigcap_{i=1}^{q} \mathbf{s}_i, \qquad \mathbf{S}_k \subseteq \{1, 2, ..., n\} \tag{2.41}$$

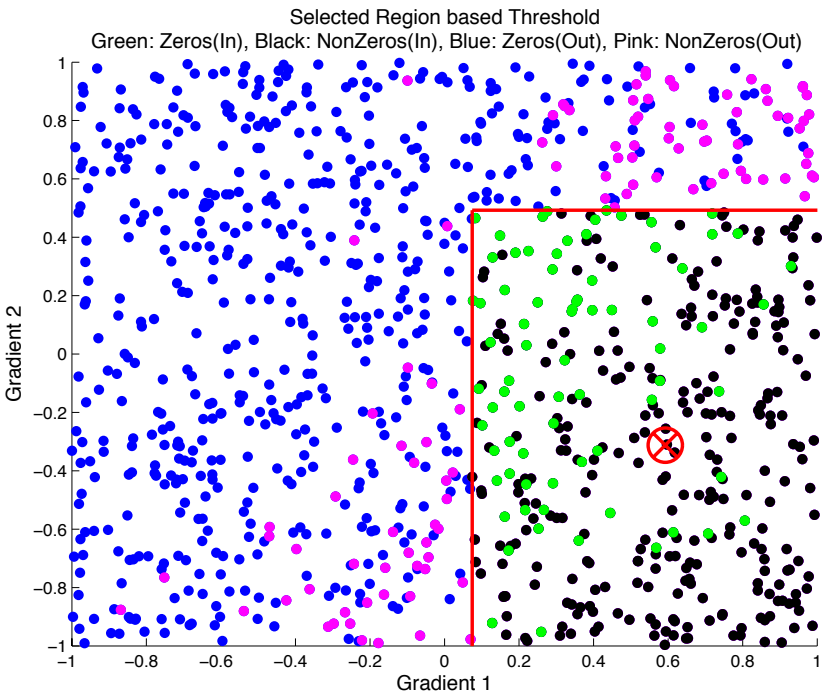

Figure 2.9: Selected regression region based on thresholding.

Figure 2.9 illustrates the new bounding region on the same data analyzed using the bounded by positive abundance method described above. The regression of this data is shown in figure 2.10.

Site selection via thresholding as described above ensures that positive abundances contribute to covariation detection, but also permits contributions of zeros towards negative covariation detection. While it is possible that an inner set of zeros between the centroid and the extreme positive abundances may be attributed to compensatory dynamics, the potential for false positives greatly outweighs detection capability in these regions. An isolated sequential series of sites with zero abundance is insufficient for the detection of negative covariation among species since it is unclear whether the outer observed positive abundances are an anomaly or some of those inner zeros may have been attributed to missed sampling opportunities. Correspondingly, evidence for negative covariation should be supported at least in part by interlaced observed positive abundances among species pairs.

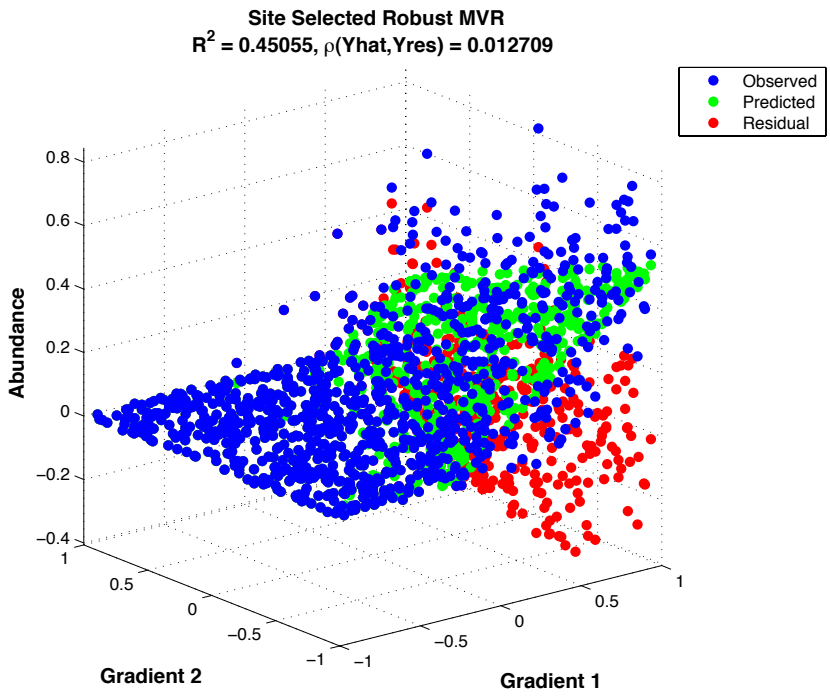

Figure 2.10: Robust regression over selected region shown in 2.9.

2.4.3 Mitigating Correlations among Predicted & Residual Responses

Based on the discussions in the previous section correlation among the predicted and residual responses have been found to be an issue for subsequent residual analyses. Several alternative approaches to standard multivariate regression have been discussed, but it is unclear which methods would be most suited to mitigate residual artifacts. The following experiment has been devised to gain insight into which potential methods produce the least correlation among the predicted and residual responses.

Experimental Setup

A total of 384 experiments have been evaluated over a range of parameters typically encountered with real ecological datasets. The amount of sites, abiotic factors and noise in the datasets have been varied with 100 iterations computed for each set of particular variables. Table 2.1 summarizes the potential methods discussed in this chapter. Regression methods evaluated include: Standard multivariate regression, Tobit regression, Standard regression bounded by positive abundance, Standard

Site Selection Methods	Regression Methods	Predicted Responses
All Sites Included	Standard MVR	All $\hat{\mathbf{y}}_k$
Positive Abundance Boundary	Tobit Regression	$\hat{\mathbf{y}}_k > 0$
Abundance Threshold Boundary	Robust Regression	

Table 2.1: Methods investigated in the analysis of species residual responses. There are 18 distinct possible configurations for consideration.

regression bounded by abundance threshold, Robust regression bounded by abundance threshold and Tobit regression bounded by abundance threshold. For each method the correlation among the predicted and residual responses is averaged over all twenty species within the community.

Results

Figure 2.11 illustrates the performance over varying the number of sites included in the community analysis. The average correlation given is the average of the correlations swept over the four gradient and the twelve noise experiments. Figure 2.12 illustrates the performance over varying the amount of noise included in the community. The average correlation is the average of the correlations swept over the four gradient and eight site experiments. Unsurprisingly, standard multivariate regression performed poorest. The robust regression bounded by abundance threshold performed the best closely followed by Tobit regression bounded by abundance threshold. Both of these methods are considered in the sensitivity analyses presented in chapter 4. It is shown that robust regression bounded by abundance threshold does indeed perform the best in terms of type I error rate.

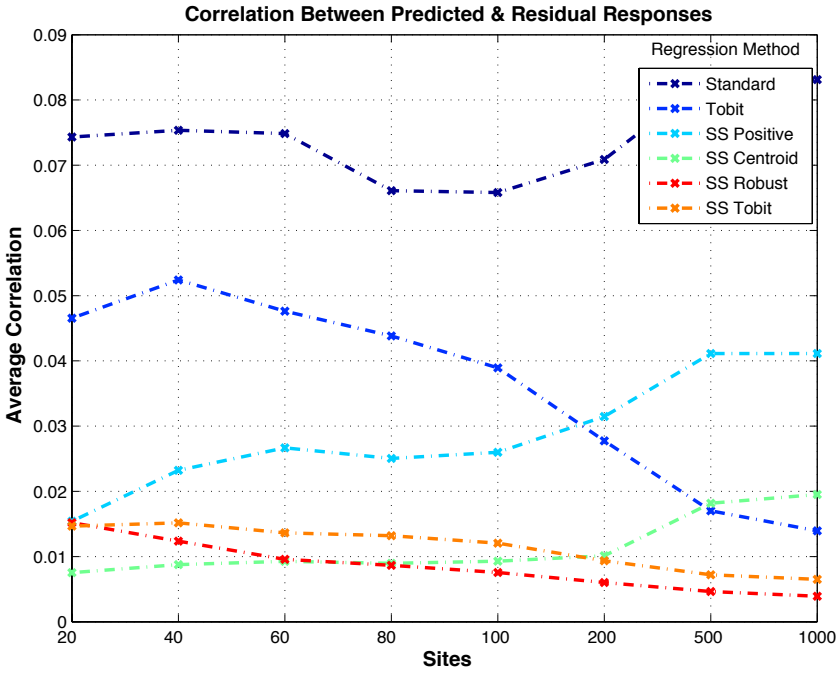

Figure 2.11: Comparing correlation between predicted and residual responses for several regression methods over varying number of sites included in the simulation.

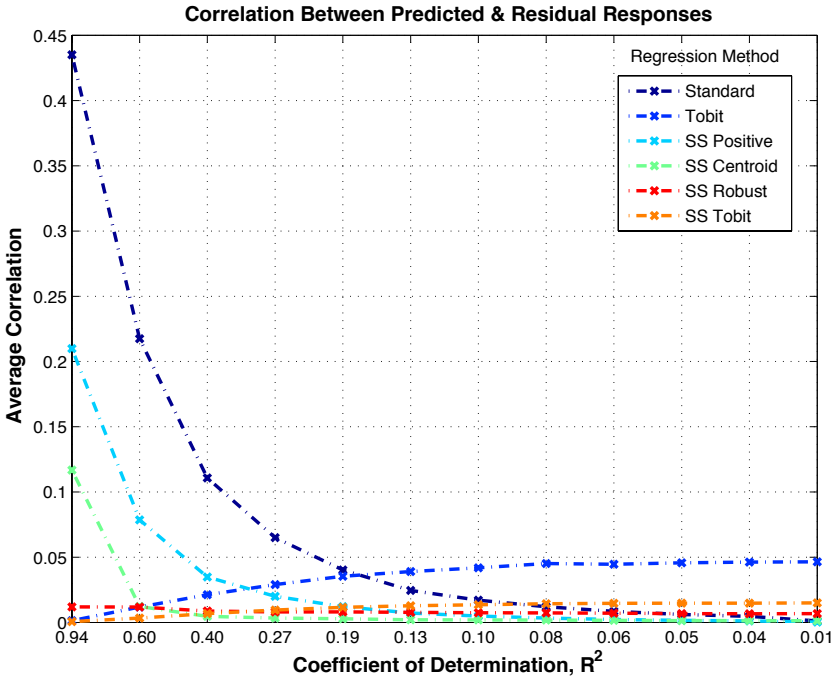

Figure 2.12: Comparing correlation between predicted and residual responses for several regression methods over varying values of the coefficient of determination.

Chapter 3: Biotic Interactions

3.1 Introduction

It is well known in ecology that both abiotic and biotic factors structure communities. This chapter focuses on the detection of biotic interactions among organisms. Methods for detection primarily rely on assessing patterns of occurrence across communities of organisms and are generally formulated within a hypothesis testing framework. Classical tests of hypotheses such as the Neyman-Pearson lemma have been explored in the literature, but have widely been replaced by a non-parametric family of methods known as null models due to difficulties parameterizing species occurrence distributions. Null models and their suitability for detecting biotic covariation will be the primary focus of this chapter.

This chapter begins with a background discussion on hypothesis testing for biotic interactions as employed in the ecology literature. Null models along with their requisite test statistics are explored for the purposes of detecting statistically significant biotic interactions across communities of organisms. Potential suitable methods are identified and described. A novel method for quantifying biotic interactions using a pairwise interaction model is developed and discussed.

3.2 Background

Null models condense ecological theory about species interactions into a prediction about the structure of a community data matrix. They involve hypothesis testing procedures developed to overcome difficulties in specifying the distributions of species responses under the influence of biotic interactions with other co-habitating species [6, 8, 29, 30, 31, 32]. Null models represent a parsimonious approach to determine if observed evidence of a predetermined pattern indicative of interactions is more extreme than expected by chance [7].

A null model consists of a randomization algorithm that serves as a rule-based model to randomize an observed community data matrix with respect to some ecological process of interest [33].

Typically, hundreds of randomized matrices are generated. Each matrix is characterized by a carefully selected index (i.e. test statistic) that is designed to quantify the pattern of interest within the community [34]. The tail probability (i.e. p-value) of the index of the observed matrix is compared to the distribution of indices computed from the randomized matrices. A decision is made regarding the outcome of the null hypothesis for the observed matrix, typically at the 5% significance level. The selection of an appropriate null model and test statistic is non-trivial and must be validated through empirical experimentation [32]. Most analyses have focused on presence and absence community data, but more recently a shift towards incorporating species traits such as abundance has occurred [8].

A summary of the general procedure for using a null model for a given application has been described in detail [6, 35]. The steps are as follows:

1. Condense ecological predictions about the structure of a community data matrix into a simple hypothesis test.
2. Define a test statistic, T that collapses a matrix down to a single number that captures the structure of interest.
3. Measure T_{obs} for the observed data matrix.
4. Randomize the observed data matrix with the null model and compute T_{sim} for the randomized matrix.
5. Repeat the step 4 numerous times (e.g. 1000) to generate a distribution for T_{sim}.
6. Compute the tail probability by finding the number of T_{sim} with a value equal or greater to T_{obs}. Divide by the number of repetitions in step 5. Reject the null hypothesis if less than 0.05.

3.2.1 Randomization Scheme

A null model specifies a randomization scheme. The observations of sites and species within a community data matrix are randomized according to a design rule derived from theoretical ecology.

Many rules have been discussed and implemented in practice, but a single rule consensus is unavailable due to differing requirements among applications of study [6, 8]. Rules range from liberal where there is an equal probability of including a cell within the matrix for randomization to conservative where row and/or column sums must be retained. There is usually an error tradeoff with liberal rules having higher Type I error rates and more conservative rules having higher Type II errors. Sensitivity analyses using simulated data are recommended for determining the performance of a chosen null model. Generally, more conservative rules featuring lower Type I error rates are preferred [32].

3.2.2 Test Statistic

Null models require a test statistic to quantify the pattern of interest within each matrix. This statistic collapses the entire matrix into a single number so that the statistic of the observed matrix may be compared with the distribution of statistics of the randomized matrices. An example of a statistic for testing hypothesis on species segregation includes computing the number of species pairs forming perfect "checkerboards" within the matrix [36]. Regardless of the statistic chosen it must be carefully selected so that it captures the pattern of interest in the hypothesis test while rejecting others. Many statistics are computed over all pairs of species within the community and then averaged together [32].

3.2.3 Performance Evaluation

Unlike parametric hypothesis tests the null model randomization scheme and the test statistic may be selected independently to provide the most appropriate test for a given hypothesis about the community. The performance of the chosen combination is generally unknown, however, and must be vetted via a sensitivity analysis to assess the type I and type II error rates. This is accomplished using Monte Carlo methods to generate sets of artificial test matrices obeying either the null or alternative hypothesis. The sensitivity analysis cannot be performed on real data since the amount of randomness of the underlying structure, if present, is unavailable [32]. The artificial data allows the control of signal and noise within the test matrices to obtain an accurate assessment of the true

performance of the null model under investigation. Once an acceptable null model is chosen based on its performance in the sensitivity analysis it may be used in practice on real data.

3.2.4 Determining Number of Repetitions

A null distribution must be constructed for the hypothesis test via randomizations of the observed data. It is infeasible to generate the complete distribution consisting of all possible permutations of the data. For example, there are 10! possible randomizations for a dataset with 10 observations, per species. A dataset with 20 species would therefore have $10!^{20}$ unique randomizations. This number is so large that it is greater than the number of atoms in the entire universe. Therefore, a subset of the set of unique randomizations must be used to construct the null distribution. Dwass first looked into this problem followed by Marriott, then Manly [37, 38, 34].

A test that involves sampling a randomization distribution is exact in the sense that using a $100\alpha\%$ level of significance has a probability of α or less of giving a significant result when the null hypothesis is true. Dwass constructed a one-sided test with observed data and a test statistic u_0 to show that a randomization test is exact even the the randomization distribution is sampled. Here a test statistic that is only a function of the relative order of the data is used. N values of u are generated by randomization of the data and the null hypothesis u_0 is a random value from this distribution. If u_0 is one of the largest m, equation (3.1), when N+1 values are ordered from smallest to largest then it is deemed significant. The probability of rejecting the null hypothesis is described in equation (3.2) where N is the number of randomizations, m is the number of significant randomizations, α is the level of significance of the test and r is the exact position of a significant randomization within the ordered distribution.

$$m = \alpha(N + 1) \tag{3.1}$$

$$P = \sum_{r=0}^{m-1} \binom{N}{r} \alpha^r (1 - \alpha)^{(N-r)} \tag{3.2}$$

		p				
$\mathbf{m}^{\alpha}$	**N+1**	**0.100**	**0.075**	**0.050**	**0.025**	**0.010**
1	20	0.135	0.227	0.377	0.618	0.826
2	40	0.088	0.199	0.413	0.745	0.942
5	100	0.025	0.128	0.445	0.897	0.997
10	200	0.004	0.065	0.461	0.971	1.000
50	1000	0.000	0.001	0.483	1.000	1.000

Table 3.1: Probability of rejecting the null hypothesis given a null distribution that has been sampled from the total set of unique possible randomizations. N+1 is the number of randomizations of the data used to construct the null distribution and m is the number of randomizations that are significant given N+1. **p** is the true p-value. This table assumes $\alpha = 0.05$. For example, with 200 randomizations there is a 0.065 probability of rejecting the null hypothesis with a p-value of 0.075. Ideally, the probability should be zero since the p-value is greater than α. Increasing the number of randomizations to 1000 decreases the probability to 0.001 as desired.

Table 3.1, as constructed by [34], shows the probabilities of rejecting the null hypothesis for various values of N ranging from 20 through 1000, given a significance level of 5% and some true underlying p-value. For example, at 200 repetitions (N+1) there is a 0.065 probability of rejecting the null hypothesis with a p-value of 0.075. Ideally, this value should be zero since we shouldn't reject the null hypothesis at this p-value. Increasing the number of repetitions N decreases the probability to 0.001 as desired. On the other hand there is a probability of 0.971 that the null hypothesis will be rejected for an observed p-value of 0.025 with 200 repetitions. Ideally, this value should be 1 since the true p-value is less than 0.05. Increasing the repetitions to 1000 increases the probability to 1.000. Based on this discussion a large number of randomizations is not necessary. At a 5% significance level 1000 repetitions is appropriate, but as few as 200 may be used in cases with intensive calculations.

3.3 Null Models for Biotic Interactions

Residual responses obtained from the multivariate regression described in chapter 2 should be void of variation explained by the environmental factors included in the regression. Interactions inferred among the residuals (i.e. covariation) may therefore be attributed to biotic interactions assuming that all relevant environmental factors were considered in the regression. The null hypothesis, H_0 on the residuals is constructed such that it is assumed there is an absence of significant covariation

Randomization Schemes
1. Permute All Sites
2. Permute Sites Satisfying $\mathbf{y}_k > 0$
3. Permutate Sites Satisfying $\hat{\mathbf{y}}_k > 0$
4. Permute Sites Selected in $\mathbf{S}_k$

Table 3.2: Randomization schemes considered for null model selection.

among pairs of species. The hypothesis test is designed to assess covariation among all pairs of species within the entire community rather than particular species pairs.

Pairwise models are not uncommon among null model implementations since corrections for multiple comparisons (e.g. Bonferroni) prohibitively decrease the significance threshold for detection [39]. For example, at a 5% significance level 100 species produces 4950 pairs (i.e. $\frac{p(p-1)}{2}$), thus requiring a p-value less than ~ 0.00001 to reject the null hypothesis. Since the number of species pairs grows exponentially with the total number of species in the community it is infeasible to test every pair for significant covariation. Considering all pairs simultaneously within the community circumvents this issue, however. An effective strategy therefore is to consider the entire community matrix initially and then follow-up with pairwise analyses using adjusted p-values on judiciously selected species pairs [32].

3.3.1 Randomization Schemes

Several null model randomizations were considered for the present method as enumerated in table 3.2. Each repetition always starts with the original residual response vector, $\mathbf{y}_{res_k}$ prior to randomization. Each species response is randomized independently of the others and sites are selected with uniform probability without replacement. The first scheme, equation (3.3) randomly permutes the abundances across all sites. The second scheme, equation (3.4) randomly permutes the abundances across sites for which a positive abundance was observed for the particular species being randomized. The third, equation (3.5) randomly permutes the abundances across sites for which there was a positive predicted abundance. The fourth scheme, equation (3.6) randomly permutes the abundances across sites which were selected using a site selection method described in chapter 2

Test Statistics	Weights
$\sigma(\mathbf{y}_{res_k}, \mathbf{y}_{res_z})^2$	No Weight
$\lvert\sigma(\mathbf{y}_{res_k}, \mathbf{y}_{res_z})\rvert$	Number of Sites
$\rho(\mathbf{y}_{res_k}, \mathbf{y}_{res_z})^2$	$\lvert\sigma(\mathbf{y}_{res_k}, \mathbf{y}_{res_z})\rvert$
$\lvert\rho(\mathbf{y}_{res_k}, \mathbf{y}_{res_z})\rvert$	$\lvert\rho(\mathbf{y}_{res_k}, \mathbf{y}_{res_z})\rvert$

Table 3.3: Test statistics and weights considered in null model selection.

$$\mathbf{y}^*_{res_k} \sim \text{random permutation of } \mathbf{y}_{res_k} \text{ over all sites} \tag{3.3}$$

$$\mathbf{y}^*_{res_k} \sim \text{random permutation of } \mathbf{y}_{res_k} \text{ over sites with with } \mathbf{y}_k > 0 \tag{3.4}$$

$$\mathbf{y}^*_{res_k} \sim \text{random permutation of } \mathbf{y}_{res_k} \text{ over sites with } \hat{\mathbf{y}}_k > 0 \tag{3.5}$$

$$\mathbf{y}^*_{res_k} \sim \text{random permutation of } \mathbf{y}_{res_k} \text{ over sites in } \mathbf{S}_k \tag{3.6}$$

3.3.2 Test Statistics

Once each response within the residual matrix has been permuted a test statistic is calculated over the entire randomized matrix. It is computed across all species pairs using only those mutual sites within the pair that have been randomized, equation (3.7). This process repeats for multiple repetitions (e.g. 1000), equation (3.8), each on a new randomized matrix to form a null distribution for the significance test.

$$Sites = \mathbf{S}^*_k \bigcap \mathbf{S}^*_z \tag{3.7}$$

$$\text{rep} = 1, 2, \ldots, \text{repetitions (e.g. 1000)} \tag{3.8}$$

Several test statistics have been developed and considered in this thesis for assessing overall community covariation as listed in table 3.3. Each of them are computed among all species pairs in the community. These include squared covariation (equation (3.9)), squared correlation (equation (3.11)), absolute covariation (equation (3.12)) and absolute correlation (equation (3.13)).

$$I_{rep} = \frac{1}{W} \sum_{k=1}^{p-1} \sum_{z=k+1}^{p} w_{zk} \, \sigma(\mathbf{y}_k^*(Sites), \mathbf{y}_z^*(Sites))^2 \tag{3.9}$$

$$W = \sum_{k=1}^{p-1} \sum_{z=k+1}^{p} w_{zk} \tag{3.10}$$

$$I_{rep} = \frac{1}{W} \sum_{k=1}^{p-1} \sum_{z=k+1}^{p} w_{zk} \, \rho(\mathbf{y}_k^*(Sites), \mathbf{y}_z^*(Sites))^2 \tag{3.11}$$

$$I_{rep} = \frac{1}{W} \sum_{k=1}^{p-1} \sum_{z=k+1}^{p} w_{zk} \, |\sigma(\mathbf{y}_k^*(Sites), \mathbf{y}_z^*(Sites))| \tag{3.12}$$

$$I_{rep} = \frac{1}{W} \sum_{k=1}^{p-1} \sum_{z=k+1}^{p} w_{zk} \, |\rho(\mathbf{y}_k^*(Sites), \mathbf{y}_z^*(Sites))| \tag{3.13}$$

Weighted averaging, equation (3.10) using weights, w_{kz} was also considered for each of these four cases as listed in table 3.3. These include no weighting, number of sites included in the calculation per species pair and the magnitude of covariation or correlation. An additional rule was also implemented where species pairs with less than a fixed number of sites were excluded from the calculation. The performance of the randomization schemes, test statistics and weights are evaluated in the sensitivity analyses of chapter 4.

3.3.3 Significance Testing

Once the null distribution has been formed the null hypothesis is assessed on the observed residual matrix by computing the number of null matrices that have test statistics equal to or greater than the test statistic of the observed matrix as described by equation (3.14). A p-value is computed and

Randomization Schemes	**Test Statistics**	**Weights**
Permute All Sites	$\sigma(\mathbf{y}_{res_k}, \mathbf{y}_{res_z})^2$	No Weight
Permute Sites Satisfying $\mathbf{y}_k > 0$	$\lvert\sigma(\mathbf{y}_{res_k}, \mathbf{y}_{res_z})\rvert$	Number of Sites
Permutate Sites Satisfying $\hat{\mathbf{y}}_k > 0$	$\rho(\mathbf{y}_{res_k}, \mathbf{y}_{res_z})^2$	$\lvert\sigma(\mathbf{y}_{res_k}, \mathbf{y}_{res_z})\rvert$
Permute Sites Selected in $\mathbf{S}_k$	$\lvert\rho(\mathbf{y}_{res_k}, \mathbf{y}_{res_z})\rvert$	$\lvert\rho(\mathbf{y}_{res_k}, \mathbf{y}_{res_z})\rvert$

Table 3.4: Potential null model configurations. An option from each column may be chosen to construct a null model.

rejects the null hypothesis if the p-value is less than 0.05 for a 5% significance level. For example, if there are 999 randomized matrices in the null distribution there must be no more than 50 test statistics that score higher than the statistic of the observed matrix to reject the null hypothesis.

$$p_{value} = \frac{\#\ of\ I_{rep} \geq I_{test}}{\#\text{ of repetitions (e.g. 1000)}} \tag{3.14}$$

$$\begin{cases} H_1 & \text{if } p_{value} < 0.05 \\ H_0 & \text{otherwise} \end{cases} \tag{3.15}$$

3.4 Simulating Biotic Interactions

Various biotic interactions could be simulated depending upon the particular study. For example, checkerboards were modeled to simulate biotic interactions such as competition in the study of presence/absence null models [6, 29]. In this thesis I introduce a pairwise biotic interaction framework that permits greater sensitivity in simulations than those employed in the mentioned studies. The interactions are superimposed on environmental responses and can be swept over an infinite range of magnitudes rather than limited to the binary values in checkerboard simulations. Checkerboards can also be simulated using this framework if desired by increasing the magnitude such that one species drives another to zero at selected sites of interest.

3.4.1 Pairwise Interaction Model

The pairwise interaction model builds upon the abiotic species simulation, equation (2.9) described in chapter 2. A matrix, M of correlated abundances are added to the abiotic responses in equation (3.16) to introduce biotic interactions.

$$\mathbf{Y} = \mathbf{XB} + \mathbf{N} + \mathbf{M} \tag{3.16}$$

For each interaction pair, c two species are chosen at random from all of the available species in the pool, equation (3.17). Species that belong to the pool do not belong to any other interaction pair and are thus free to be paired. An interaction abundance, $z_j^{(c)}$ is sampled as in equation (3.18) for each interaction pair, c at each site j.

$$\begin{aligned} a^{(c)}, b^{(c)} &\sim U(\text{pool}) \\ \text{pool} &= \{k \in \{1, 2, \ldots, p\}, k \notin a, b\} \\ c &= 0, 1, 2, \ldots, v \quad \text{interacting pairs} \end{aligned} \tag{3.17}$$

$$z_j^{(c)} \sim N(0, \sigma_M^2) \tag{3.18}$$

The interaction matrix, M is populated as in equation (3.19) such that species $a^{(c)}$ and $b^{(c)}$ have the same abundance at each site, j if the interaction Type is 1 (i.e. positive interaction). If the interaction Type is 0 (i.e. negative interaction) then the same abundance added to $a^{(c)}$ is subtracted from $b^{(c)}$. All other entries of M are zero and thus do not alter the abiotic responses previously generated.

$$M_{jk} = \begin{cases} z_j^{(c)} & \text{if } k = a^{(c)} \\ z_j^{(c)} & \text{if } k = b^{(c)} \ \& \ \text{Type} = 1 \ \text{(Positive)} \\ -z_j^{(c)} & \text{if } k = b^{(c)} \ \& \ \text{Type} = 0 \ \text{(Negative)} \\ 0 & \text{otherwise} \end{cases} \tag{3.19}$$

All negative abundances are truncated at zero to impose the realistic constraint of species count data, equation (3.20).

$$\mathbf{Y} = \mathbf{Y}(< 0) = 0 \tag{3.20}$$

3.4.2 Quantifying Community Covariation

Once significant biotic interaction has been found in a community data matrix the type of covariation may be assessed. A procedure similar to the null model as discussed above is used to determine whether the community has significant positive or negative covariation. The test statistic in equation (3.21) is employed on randomizations of the residual response matrix to generate a null distribution.

$$I = \frac{1}{p(p-1)/2} \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \sigma(\mathbf{y}_{\mathbf{res}i}, \mathbf{y}_{\mathbf{res}j}) \tag{3.21}$$

Each test statistic is indexed by the respective repetition, I_{rep}. Generally, 1000 repetitions are computed to generate the null distribution. The test statistic, I_{test}, is then computed on the residual response matrix. A p-value is computed for significant positive (equation (3.22)) and negative (equation (3.23)) interactions. The level of significance is generally set at $\alpha = 0.05$. If $p_{positive}$ is less than or equal to α then the community is deemed to have significant positive interactions. Alternatively, if $p_{negative}$ is less than or equal to α the community is deemed to have significant negative interactions.

$$p_{positive} = \frac{\#\ I_{rep} \geq I_{test}}{\#\ \text{of repetitions}} \tag{3.22}$$

$$p_{negative} = \frac{\#\ I_{rep} \leq I_{test}}{\#\ \text{of repetitions}} \tag{3.23}$$

Chapter 4: Sensitivity Analyses

4.1 Introduction

The goal of this chapter is to explore the performance of the methods discussed in chapters 2 and 3 using Monte Carlo simulations of species response data. Chapter 2 discussed species responses to environmental factors, methods for simulating species responses to these factors and methods for modeling the responses. Chapter 3 discussed species interactions among community members, null models for testing hypotheses on the presence of these interactions and a method for simulating species interactions. This chapter draws upon the previous two by investigating the performance of environmental factor models coupled with null models for hypothesis testing of species interactions. The benefits of such a hybrid approach include hypothesis testing of species interactions after accounting for environmental variation, testing of more subtle interactions than present methods permit and partitioning of variation explained by environmental factors and biotic interactions in community data.

This chapter begins with the motivation behind using simulated species responses in the subsequent experiments in this chapter. This is followed by an overview of the hybrid methods explored in these experiments. The hybrid methods are comprised of combinations of the methods discussed in chapters 2 and 3. Experiments exploring the Type I error performance (false positive rate) are discussed first followed by experiments on true positive rate (i.e. power, sensitivity, 1-Type II error) performance. The latter set of experiments are conducted only on those exhibiting favorable Type I errors. The chapter concludes with a summary of the experiments and the identification of a superior hybrid method.

4.1.1 Motivation

Traditional hypothesis testing such as the Neyman-Pearson (NP) lemma are well suited tests when a likelihood ratio is well defined [40]. In fact the NP is the most powerful test available given a

likelihood ratio, significance level, α and threshold η. Ecological data, however, rarely originates from a distribution from which parameters can be easily estimated [32]. Instead non-parametric approaches known as null models are employed for hypothesis testing and are more flexible than classical parametric tests [34].

Null models do not guarantee that the null and alternative hypotheses are mutually exclusive and exhaustive. Therefore the rejection of the null hypothesis with probability α does not guarantee that the alternative hypothesis is true with probability $1 - \alpha$. Thus null models need to be qualified numerically for type I and type II errors. It is desirable to conservatively optimize a null model by controlling for type I over type II errors [32]. This is because ecological data is rarely obtained via controlled experiments, but is instead observational with minimal additional information available beyond the community matrix.

Real data has unknown amounts of structure and randomness with regards to the null model hypothesis. It is important that performance tests be conducted using simulated rather than empirical data [6]. Simulations permit control of signals and noise in the data and allow for direct comparison of the performance of multiple approaches. Simulations should be parameterized (e.g. number of sites and species) by realistic conditions expected to be encountered in practice. A null model should only be employed on empirical data once it has been qualified by performance benchmarks on simulated data. The final null model chosen for a particular application should be selected based on simplicity, biological realism and performance in benchmark tests using simulated data [32].

4.1.2 Overview

Table 4.1 provides a list of all possible options considered for each portion of the hybrid method. There are 1728 possible combinations to be considered. Many options can be ruled out initially, however, so that the computational expense of the experiments can be reduced.

All experiments that do not truncate $\hat{\mathbf{Y}}$ at zero may be removed from the list since this was shown to be a major cause of correlation among predicted and residual responses in chapter 2. Standard MVR using all sites may also be removed from further consideration due to similar reasons as discussed in chapter 2. Null models considering selected sites are considered only for experiments

Site Selection, SS	Regression	$\hat{\mathbf{Y}}$	Null Model	Test Statistic	Weights
All Sites	Standard MVR	All $\hat{\mathbf{Y}}$	All Sites	$\sigma(\mathbf{y}_{res_k}, \mathbf{y}_{res_z})^2$	No Weight
Positive Bound	Robust	$\hat{\mathbf{Y}}(< 0) = 0$	All $\hat{\mathbf{Y}} > 0$	$\lvert\sigma(\mathbf{y}_{res_k}, \mathbf{y}_{res_z})\rvert$	# of Sites
Threshold Bound	Tobit		All SS	$\rho(\mathbf{y}_{res_k}, \mathbf{y}_{res_z})^2$	$\lvert\sigma(\mathbf{y}_{res_k}, \mathbf{y}_{res_z})\rvert$
				$\lvert\rho(\mathbf{y}_{res_k}, \mathbf{y}_{res_z})\rvert$	$\lvert\rho(\mathbf{y}_{res_k}, \mathbf{y}_{res_z})\rvert$

Table 4.1: Overview of options to consider for hybrid method. A filtering option that requires a minimum number of mutual sites between $\mathbf{y}_k, \mathbf{y}_z$ in computing the test statistic is also considered. All options are discussed further in chapters 2 and 3.

where sites have been selected for regression (i.e. Positive or Threshold Bound). The remaining 704 experiments are considered below in the false positive analyses.

The false positive experiments proceed by first grouping analyses by the three null model options: all sites, $\hat{\mathbf{y}}_k > 0$ and selected sites. The performance is compared over the environmental factor, sites and noise parameter sweeps respectively. The best performing test statistic and weight pair for each site selection, regression method and null model combination are then compared across the parameter sweeps to determine the best performing methods out of all combinations. Well-performing methods (i.e. Tobit regression and Threshold Bounded Robust Regression) are then scrutinized further.

4.2 False Positive Analyses (Type I Errors)

Type I errors arise when the method incorrectly classifies a community as exhibiting significant biotic interactions. Given a 5% statistical significance level, α experiments should classify no more than 5 out of 100 community matrices incorrectly. All experiments herein are repeated 100 times to estimate the significance level. An ideal method should have $\alpha \approx 0.05$.

4.2.1 Parameter Sweeps

Community data for the following experiments are simulated by the abiotic response simulations described by equations (2.1) thru (2.10) in chapter 2. A series of parameters influence the construction of the species response simulations. The following experiments evaluate type I error performance via parameter sweeps over conditions consistent of typical empirical biological data. Parameters as

Parameter	Values
# of Species	20
# of Env. Factors	1,2,3,4
# of Sample Sites	20, 40, 60, 80, 100, 200, 500, 1000
Noise Variance, σ^2_N	0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 2

Table 4.2: Parameter and value sweeps for the type I error experiments. There are a total of 384 parameter configurations evaluated for each experiment. Experiments repeated 100 times. Each null distribution is constructed using 1000 repetitions.

Noise Variance, σ^2_N	0.01	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	2
R^2_{avg}	0.94	0.60	0.40	0.27	0.19	0.13	0.10	0.08	0.06	0.05	0.04	0.01

Table 4.3: The noise variance negatively affects the fit of the abiotic response. Therefore the R^2_{avg} decreases as the noise increases. The noise is swept over all realistic values for explained abiotic variation.

configured in the simulations are summarized in table 4.2. Note that the noise is swept over all realistic values to encountered in practice as shown in 4.3.

4.2.2 Hybrid Method Comparisons

False Positives: Null Model - All Sites

Figure 4.1 shows the type I error rates for the hybrid regression and null model methods swept over the number of environmental factors included in the analysis. The rates displayed for each hybrid method are the lowest based on the twenty four possible test statistic and weight combinations explored. This figure shows the best possible performance obtainable of all combinations using all sites in the null model randomization. Figures 4.2 and 4.3 show the performance swept over the number of sites and R^2 values respectively. Note that in each of the three plots the rate at each point is averaged over all parameter sweeps. For example, in figure 4.2 the false positive rate displayed for Tobit regression with one environmental factor is the average of the sites and noise parameter sweeps.

Tobit regression with the all sites null model is the best performing hybrid method based on the analysis of figures 4.1 thru 4.3. As can be seen in figure 4.1 , the false positive rates are greater than 0.05 over 2, 3, and 4 gradient experiments. These rates are too high for a conservative hybrid method.

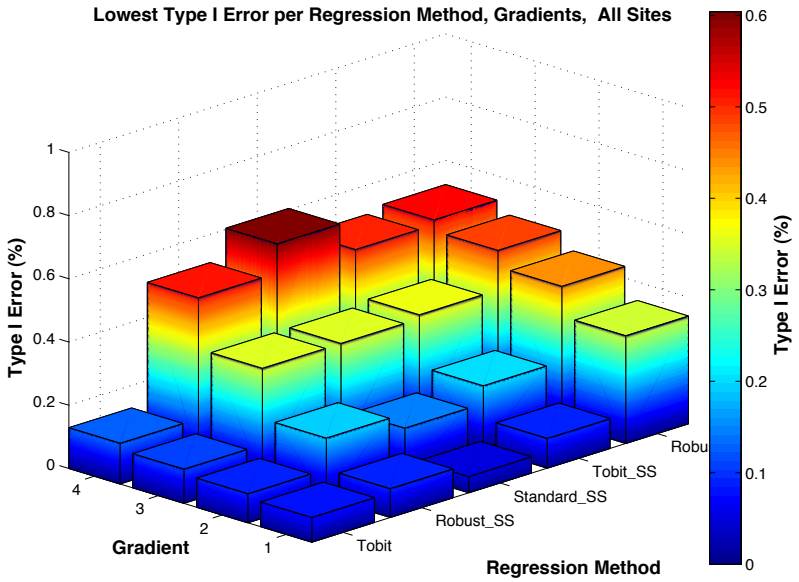

Figure 4.1: Lowest type I error performance per regression method using all sites in the null model. This figure displays the performance over the number of gradients included in the model.

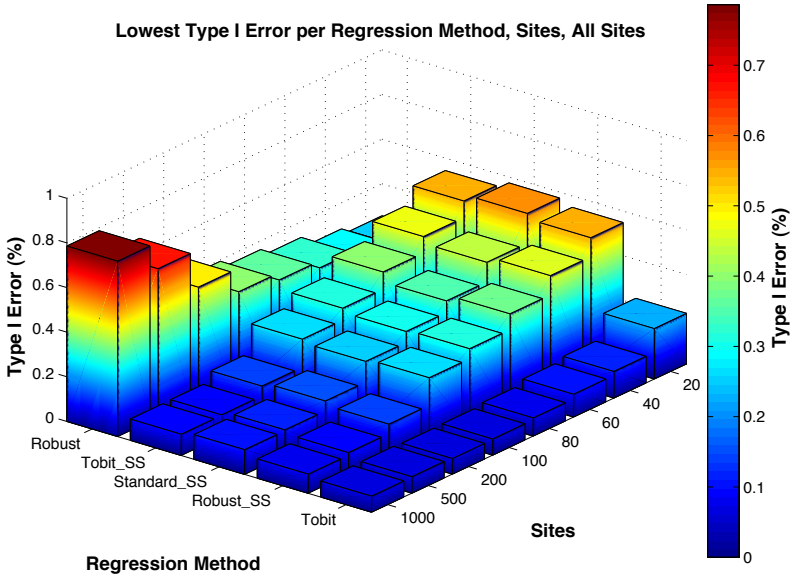

Figure 4.2: Lowest type I error performance per regression method using all sites in the null model. This figure displays the performance over the number of sites included in the model.

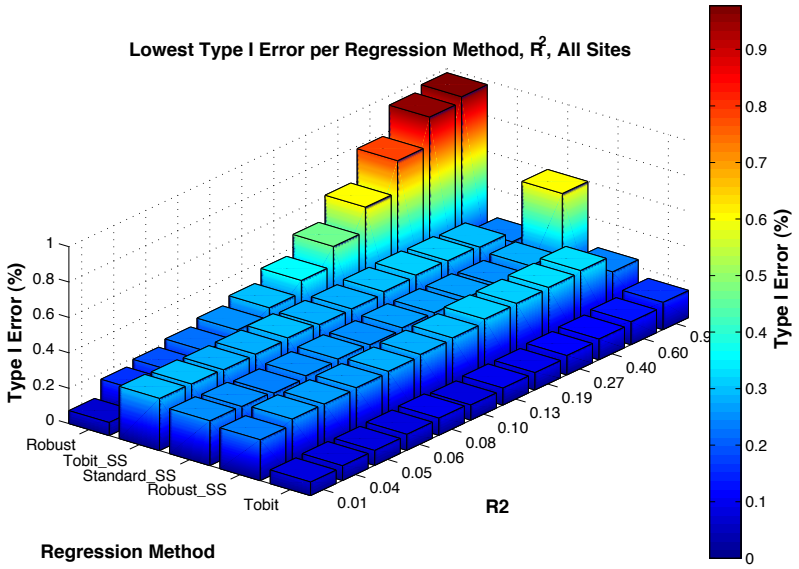

Figure 4.3: Lowest type I error performance per regression method using all sites in the null model. This figure displays the performance as a measure over the amount of noise included in the model.

False Positives: Null Model: $\hat{\mathbf{Y}} > 0$

Figure 4.4 shows the type I error rates for the hybrid regression and null model methods swept over the number of environmental factors included in the analysis. The rates displayed for each hybrid method are the lowest based on the twenty four possible test statistic and weight combinations explored. This figure shows the best possible performance obtainable of all combinations using sites with a predicted abundance greater than zero in the null model randomization. Figures 4.5 and 4.6 show the performance swept over the number of sites and R^2 values respectively.

Tobit regression with the $\hat{\mathbf{Y}} > 0$ null model is the best performing hybrid method based on the analysis of figures 4.4 thru 4.6. It appears to have a false positive rate near 0.05 over a wide range of parameters and thus may be a good candidate for a hybrid method.

False Positives: Null Model: SS Threshold

Figure 4.7 shows the type I error rates for the hybrid regression and null model methods swept over the number of environmental factors included in the analysis. The rates displayed for each hybrid method are the lowest based on the twenty four possible test statistic and weight combinations explored. This figure shows the best possible performance obtainable of all combinations using the sites selected for regression in the null model randomization. Figures 4.8 and 4.9 show the performance swept over the number of sites and R^2 values respectively.

Robust regression with the site selected null model is the best performing hybrid method based on the analysis of figures 4.1 thru 4.7. As can be seen in figure 4.9 the false positive rate is about 0.05 and below which is the intended goal.

False Positives: Comparing All Regression/Null Model Pairs

Figure 4.10 shows the type I error rates for the hybrid regression and null model methods swept over the number of environmental factors included in the analysis. The rates displayed for each hybrid method are the lowest based on the twenty four possible test statistic and weight combinations explored. This figure shows the best possible performance obtainable of all regression/null model

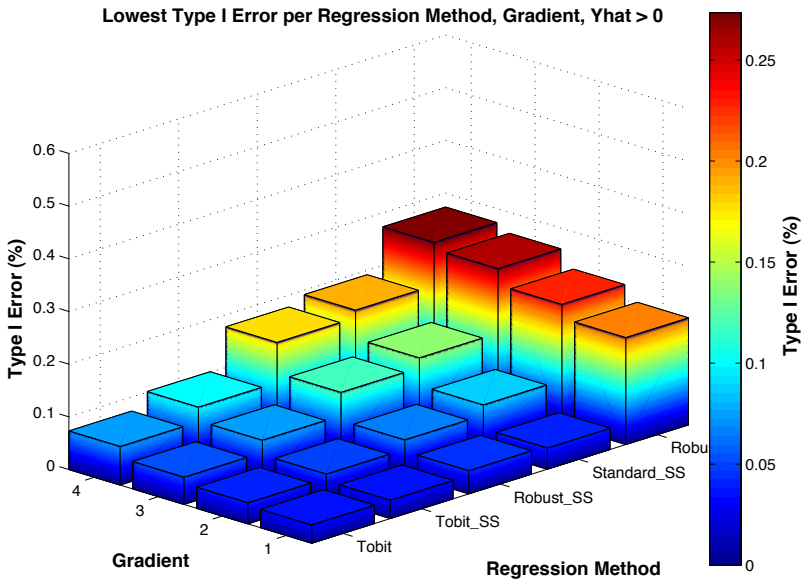

Figure 4.4: Lowest type I error performance per regression method using sites predicted with positive abundance in the null model. This figure displays the performance over the number of gradients included in the model.

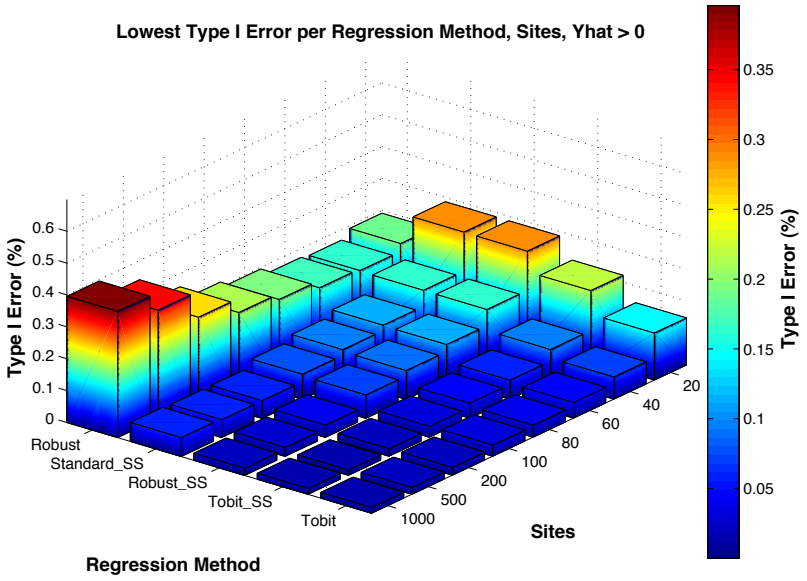

Figure 4.5: Lowest type I error performance per regression method using sites predicted with positive abundance in the null model. This figure displays the performance over the number of sites included in the model.

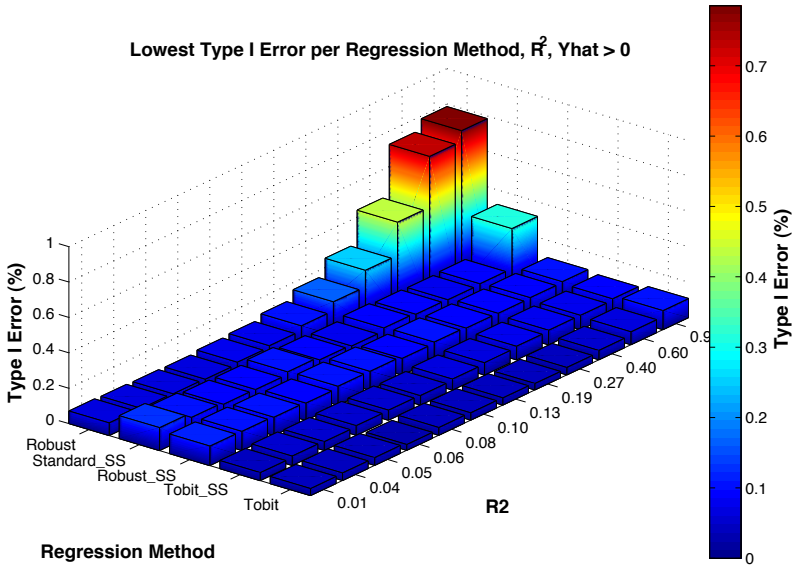

Figure 4.6: Lowest type I error performance per regression method using sites predicted with positive abundance in the null model. This figure displays the performance as a measure over the amount of noise included in the model.

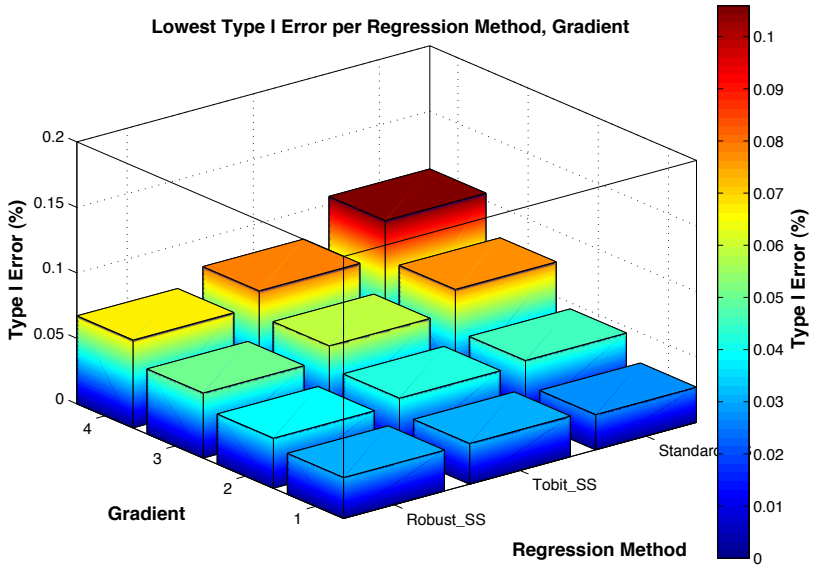

Figure 4.7: Lowest type I error performance per regression method using sites selected (SS) using the threshold method. This figure displays the performance over the number of gradients included in the model.

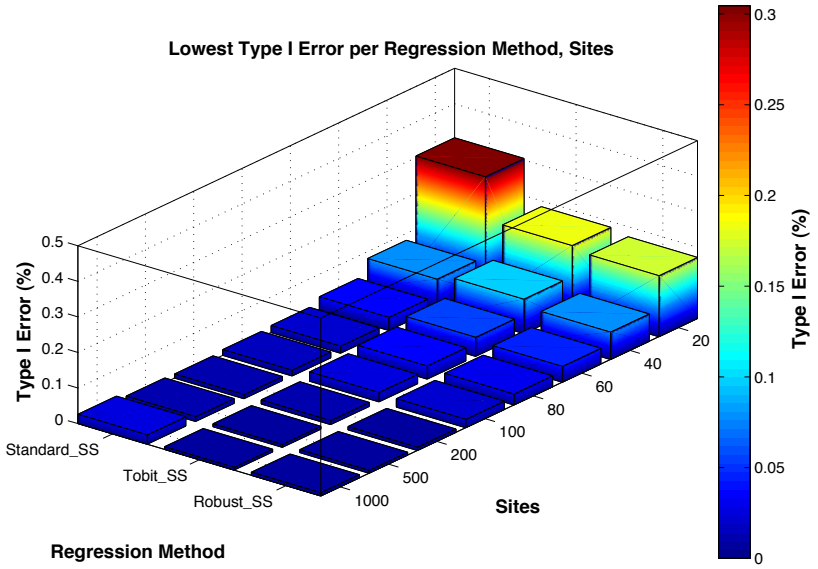

Figure 4.8: Lowest type I error performance per regression method using sites selected (SS) using the threshold method. This figure displays the performance over the number of sites included in the model.

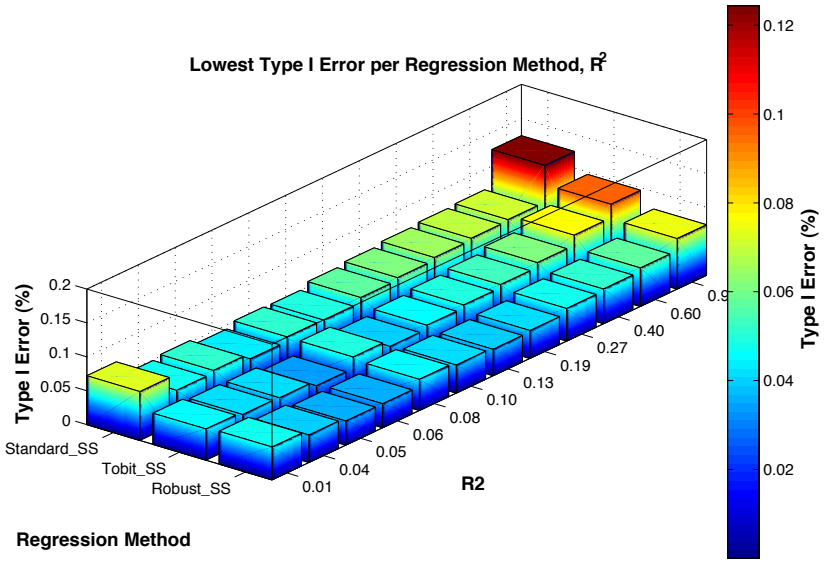

Figure 4.9: Lowest type I error performance per regression method using sites selected (SS) using the threshold method. This figure displays the performance as a measure over the amount of noise included in the model.

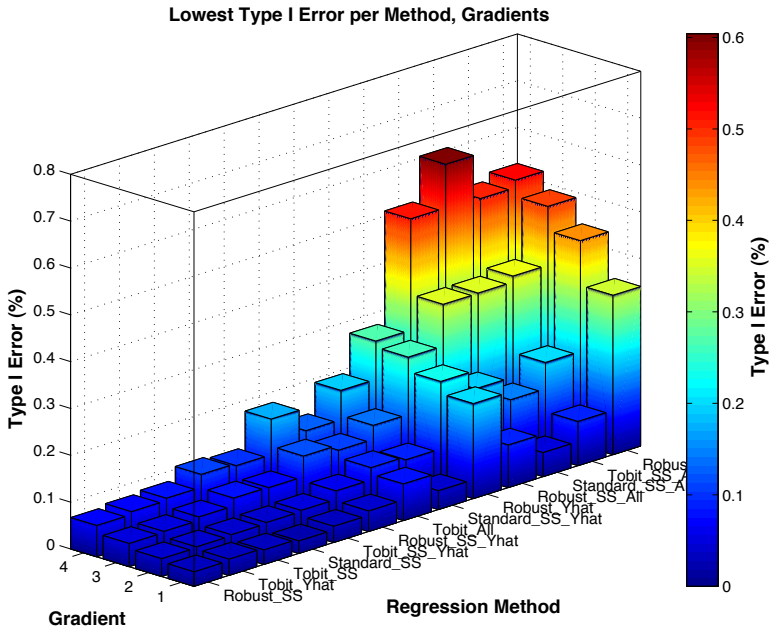

Figure 4.10: Lowest type I error performance per regression method using sites. All null models are considered in the figure. This figure displays the performance over the number of gradients included in the model.

pairs. Figures 4.11 and 4.12 shows the performance swept over the number of sites and R^2 values respectively.

Based on figures 4.10 thru 4.12, the best performing hybrid methods are site selected robust regression with a site selected null model, followed by Tobit regression with the $\hat{\mathbf{Y}} > 0$ null model and the site selected Tobit regression with site selected null model.

False Positives: SS Robust Regression/SS Null Model Test Statistics and Weights

Figure 4.13 shows the performance of SS robust regression with the SS null model using the 24 test statistic/weight combinations swept over the number of environmental factors included in the analysis. Figures 4.14 and 4.15 shows the performance swept over the number of sites and R^2 values respectively. There is a clear difference in false positive rate between the pairs 1 thru 12 and 13 thru 24. Pairs 1 thru 12 all correspond to test statistic/weight pairs involving covariance while 13 thru 24 involve correlations. The best performing test statistic is $\sigma(\mathbf{y}_{res_k}, \mathbf{y}_{res_z})^2$ with weight $|\sigma(\mathbf{y}_{res_k}, \mathbf{y}_{res_z})|$. The test statistic also includes the rule that any species pair with less than a fixed number of mutual sites are omitted from the analysis. The fixed number was selected to be 7 based on the performance tradeoff of rejecting too few/too many pairs and the false positive rate.

4.2.3 Robust Regression

Robust regression on sites selected by threshold boundary coupled with a null model using the selected sites and the test statistic $\sigma(\mathbf{y}_{res_k}, \mathbf{y}_{res_z})^2$ with weight $|\sigma(\mathbf{y}_{res_k}, \mathbf{y}_{res_z})|$ is the best performing method in terms of false positive rate based on analyses of figures 4.10 thru 4.15 in the previous section. The following set of 5 figures shows the false positive performance of this hybrid method in greater detail.

Robust Regression FP Rate Discussion

The number of sites included in the analysis, prior to the selection procedure, is the most important parameter is determining the false positive rate of detection. A minimum of 20 sites should be included based on the 20 species simulations. More than 40 sites should be included when the number of species is increased from 20 to 500. The results are plausible since the number of sites

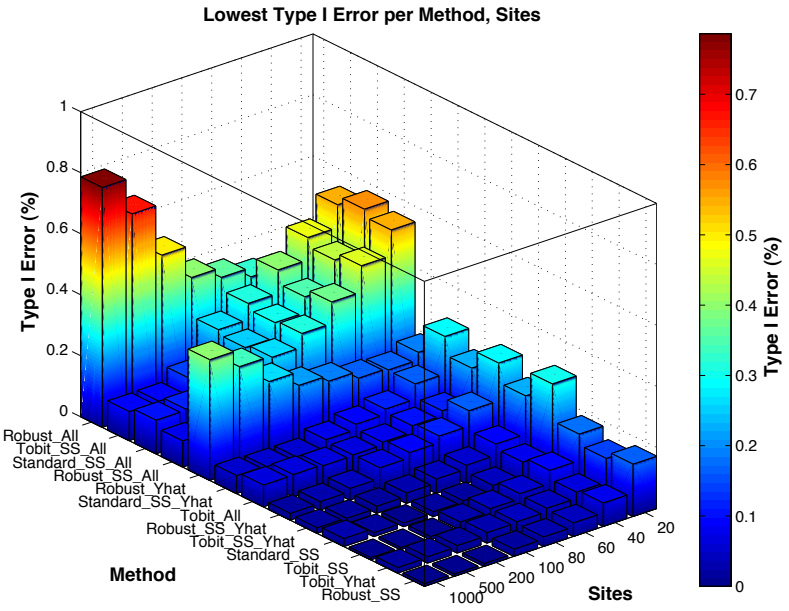

Figure 4.11: Lowest type I error performance per regression method using sites. All null models are considered in the figure. This figure displays the performance over the number of sites included in the model.

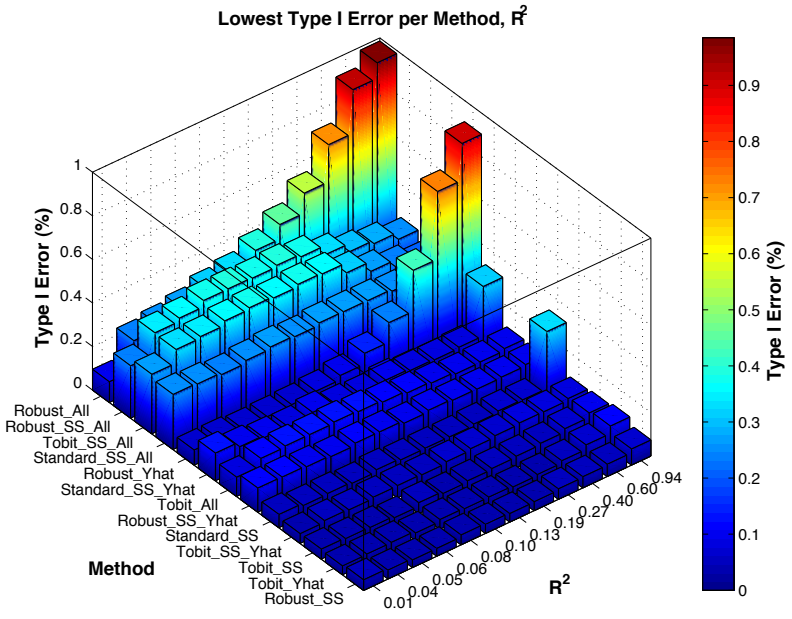

Figure 4.12: Lowest type I error performance per regression method using sites. All null models are considered in the figure. This figure displays the performance as a measure over the amount of noise included in the model.

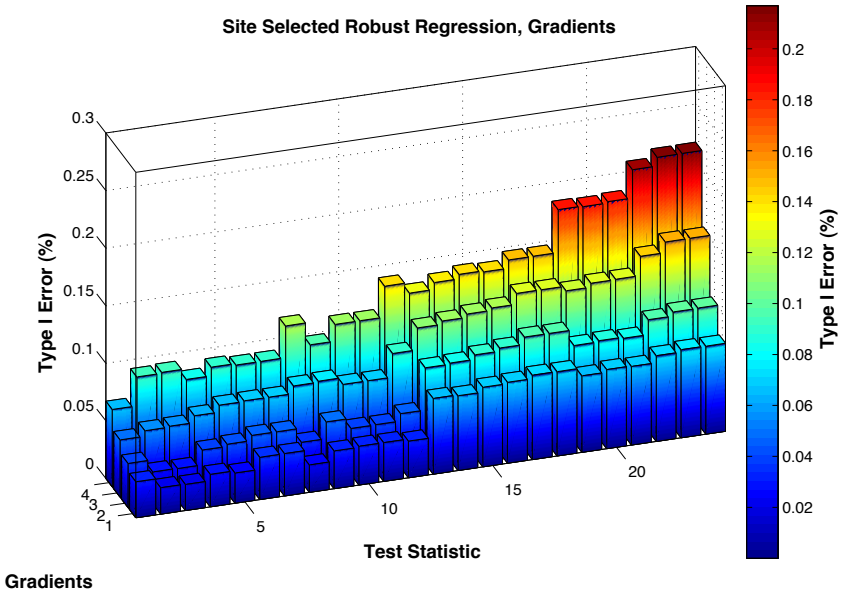

Figure 4.13: Type I error performance of robust regression with the SS threshold approach displayed over the number of environmental gradients. All test statistic/weight combinations are displayed.

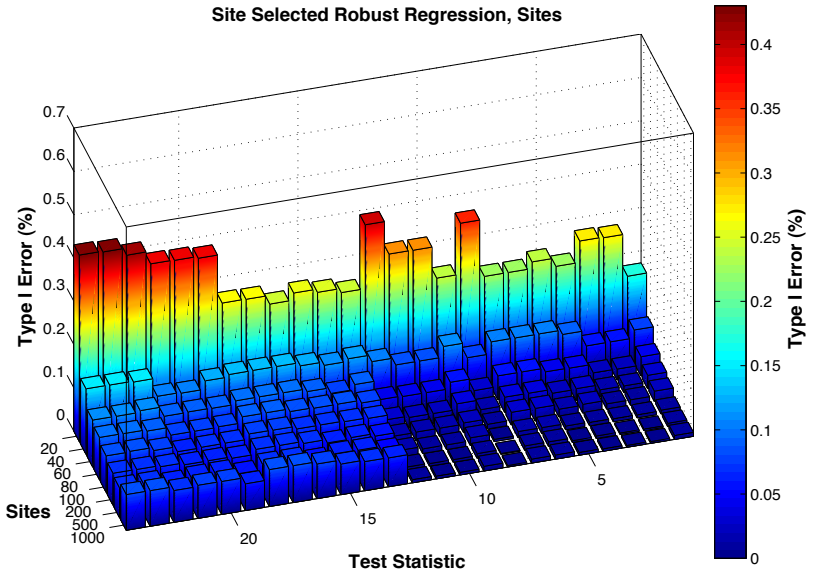

Figure 4.14: Type I error performance of robust regression with the SS threshold approach displayed over the number of sites included in the model. All test statistic/weight combinations are displayed.

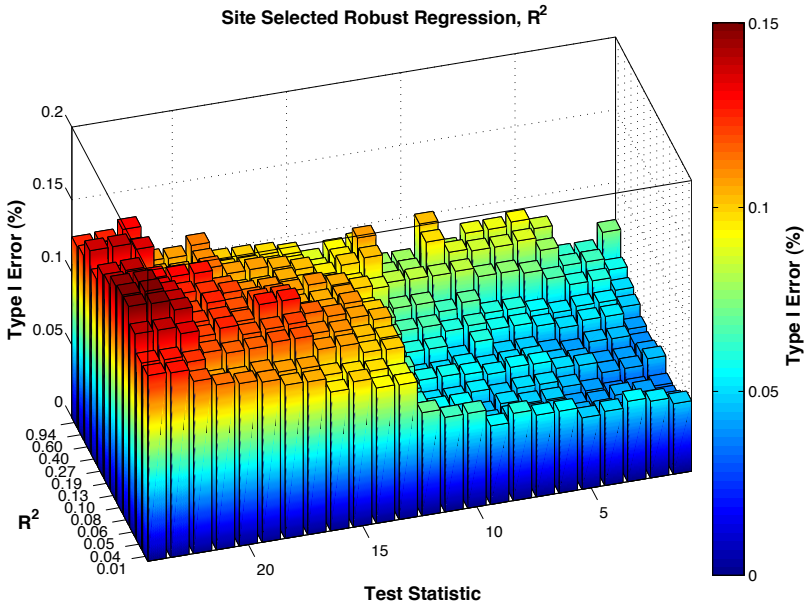

Figure 4.15: Type I error performance of robust regression with the SS threshold approach displayed over a measure of the amount of noise included in the model. All test statistic/weight combinations are displayed.

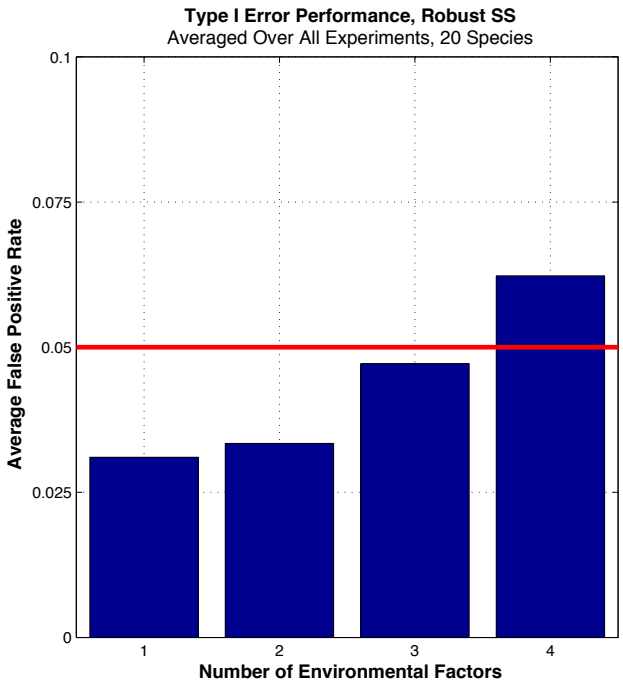

Figure 4.16: Type I error performance of SS robust regression fixed on the number of environmental factors and averaged over all experiments. The null model consisted of 1000 randomizations. This figure shows an acceptable false positive rate for use in practice.

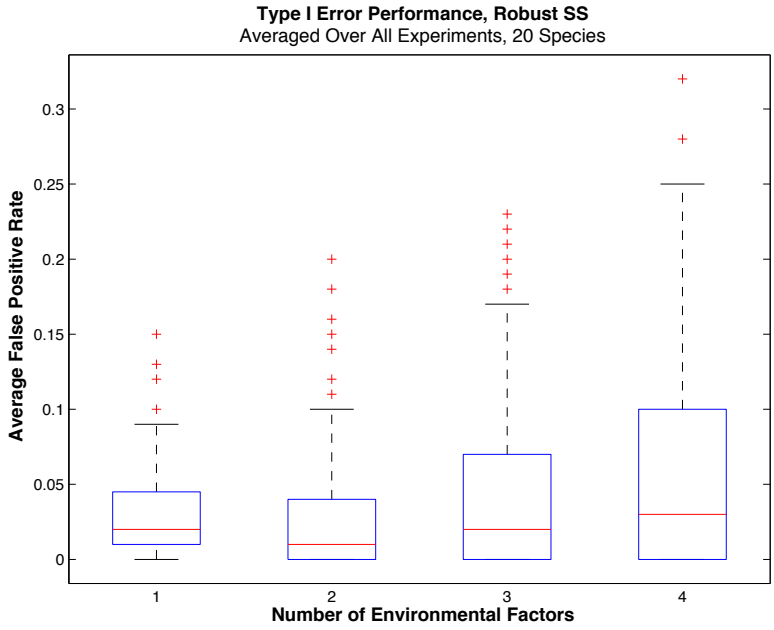

Figure 4.17: Box plot: Type I error performance of SS robust regression fixed on the number of environmental factors. The null model consisted of 1000 randomizations. This figure shows an acceptable false positive rate for use in practice.

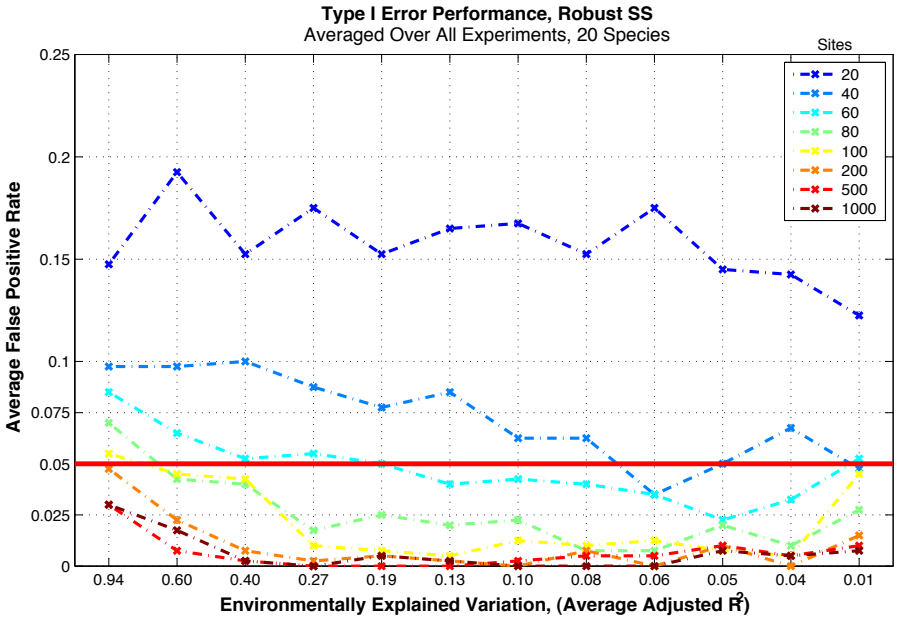

Figure 4.18: Type I error performance of SS robust regression fixed on the average R^2 value measured for the community. The null model consisted of 1000 randomizations. This figure shows an acceptable false positive rate for use in practice as long as more than 20 sites are included in the analysis.

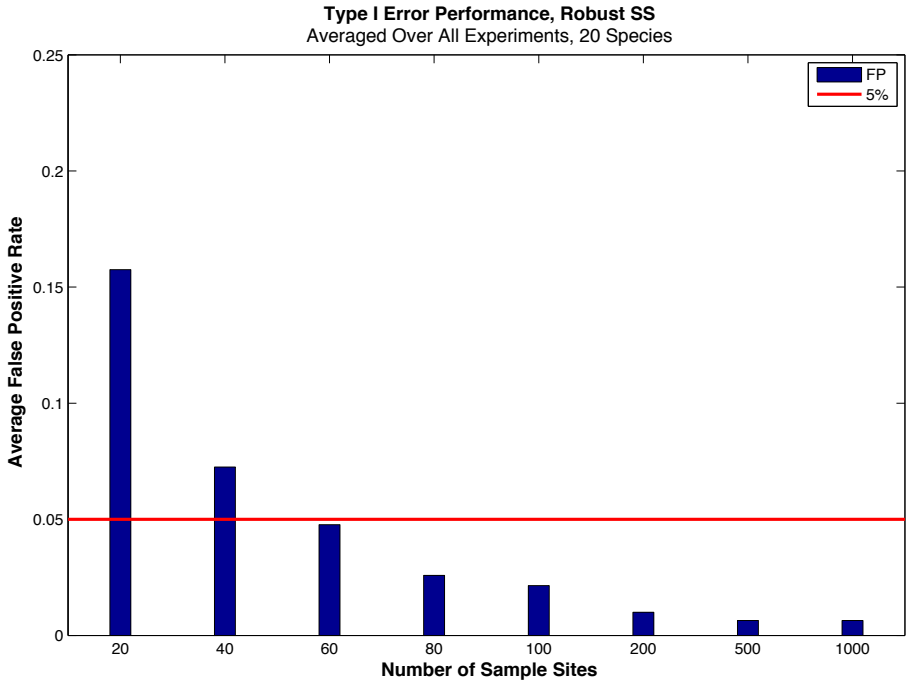

Figure 4.19: Type I error performance of SS robust regression fixed on the number of sites factors averaged over all experiments. The null model consisted of 1000 randomizations. This figure shows an acceptable false positive rate for use in practice.

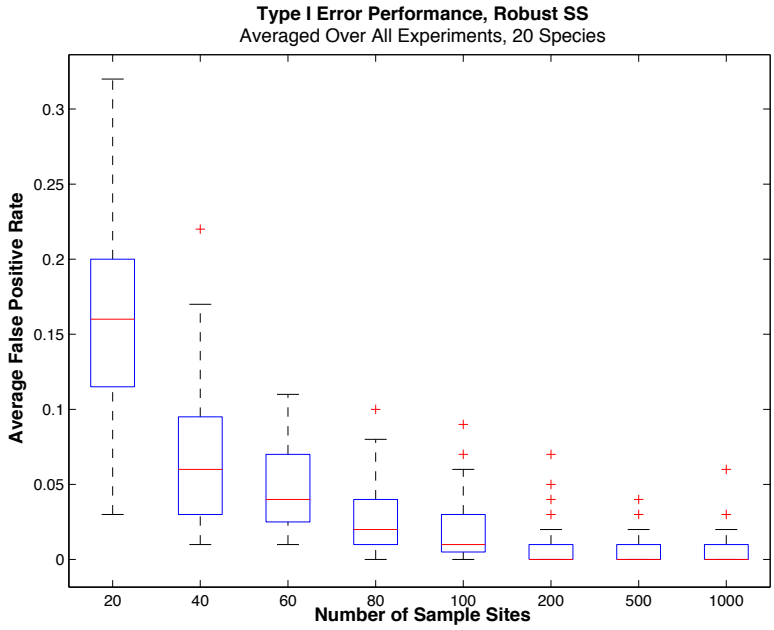

Figure 4.20: Box plot: Type I error performance of SS robust regression fixed on the number of sites. The null model consisted of 1000 randomizations. This figure shows an acceptable false positive rate for use in practice.

analyzed must be sufficiently large so that the number of possible permutations within the null model permutation universe is large compared to the number selected in the covariation hypothesis test. The number of explanatory environmental factors included in the analysis also contributes to the false positive rate, albeit much less substantially than the number of sites. The increase of Type I error with increasing environmental factors is most likely attributed to increasing model uncertainty in high dimensional space due to the curse of dimensionality. The observed increase in false positive rate is not detrimental to overall algorithm performance, however, since studies generally include no more than several quantitative environmental factors in practice.

4.2.4 Tobit Regression

Tobit regression on all sites with a null model using only sites with predicted values greater than zero (i.e. $\hat{\mathbf{Y}} > 0$) and the test statistic $\sigma(\mathbf{y}_{res_k}, \mathbf{y}_{res_z})^2$ with weight $|\sigma(\mathbf{y}_{res_k}, \mathbf{y}_{res_z})|$ is the *second* best performing method in terms of false positive rate based on analyses of figures 4.10 thru 4.15 in the previous section. The following set of 5 figures shows the false positive performance of this hybrid method in greater detail.

4.3 True Positive Analysis (Sensitivity)

A sensitivity analysis is constructed to determine the true positive detection rate of a method. It is directly related to the type II error rate (i.e. false negatives) by 1-type II error. Due to an acceptable false positive rate, the robust regression method as discussed above will be explored in this section for detection performance of species interactions. The advantage of Robust regression's performance over Tobit regression can be clearly seen in figure 4.26. Species interactions are quantified in terms of the covariance between pairs of species within the community as discussed in chapter 3.

There are considerably more parameters to sweep in the evaluation of sensitivity. In addition to the number of environmental factors, sites and noise parameters considered in the false positive analysis, additional parameters must be included to account for the number of covarying species pairs within the community matrix, the type and amount of covariation. The amount of covariation may be though of as a signal-to-noise ratio so it is selected with respect to the noise parameter. For

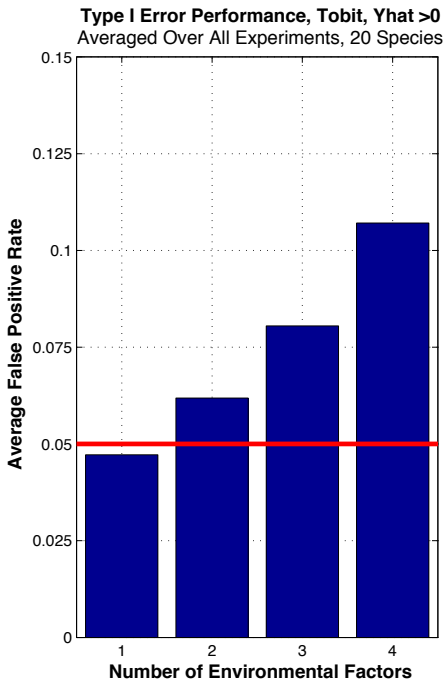

Figure 4.21: Type I error performance of Tobit regression fixed on the number of environmental factors and averaged over all experiments. The null model consisted of 1000 randomizations. The false positive rate is higher than SS robust regression.

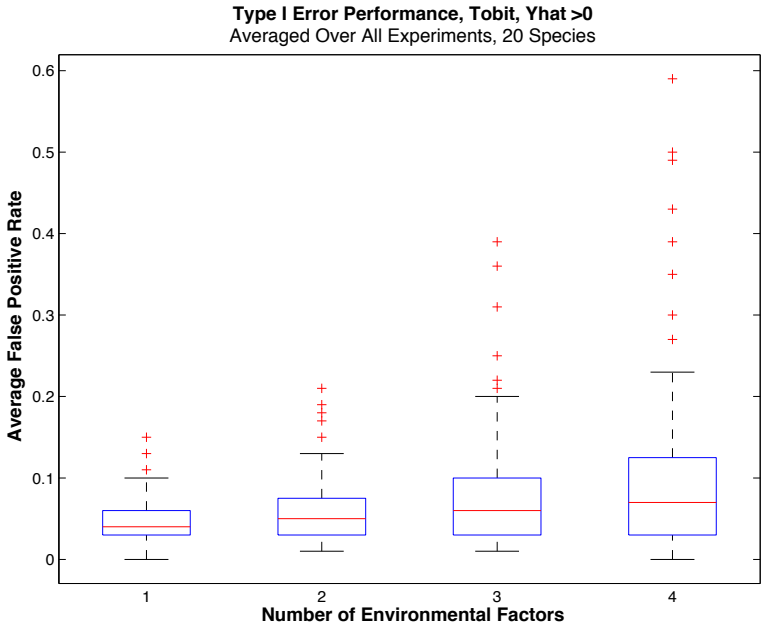

Figure 4.22: Box plot: Type I error performance of Tobit regression fixed on the number of environmental factors. The null model consisted of 1000 randomizations. The false positive rate is higher than SS robust regression.

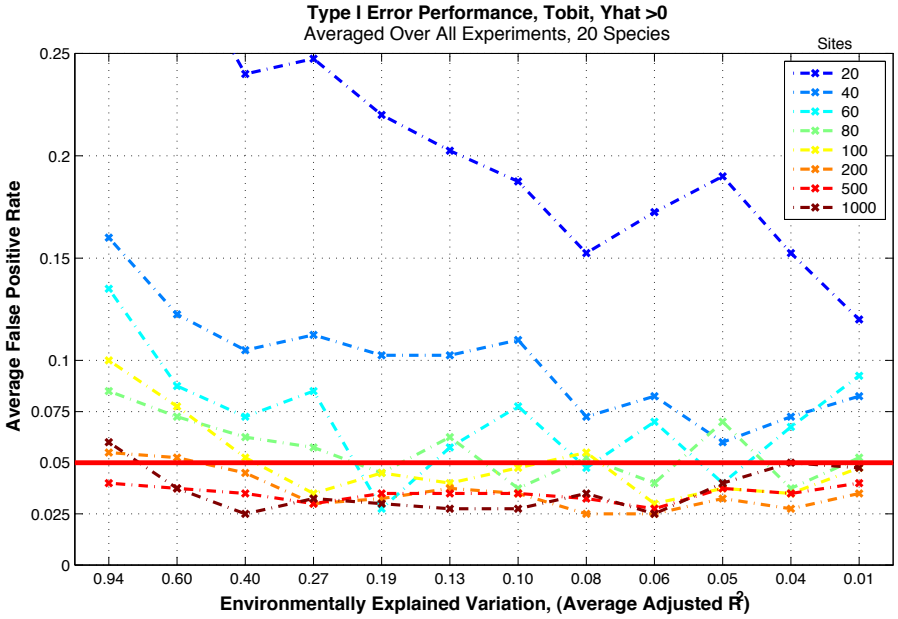

Figure 4.23: Type I error performance of Tobit regression fixed on the average R^2 value measured for the community. The null model consisted of 1000 randomizations. The false positive rate is higher than SS robust regression.

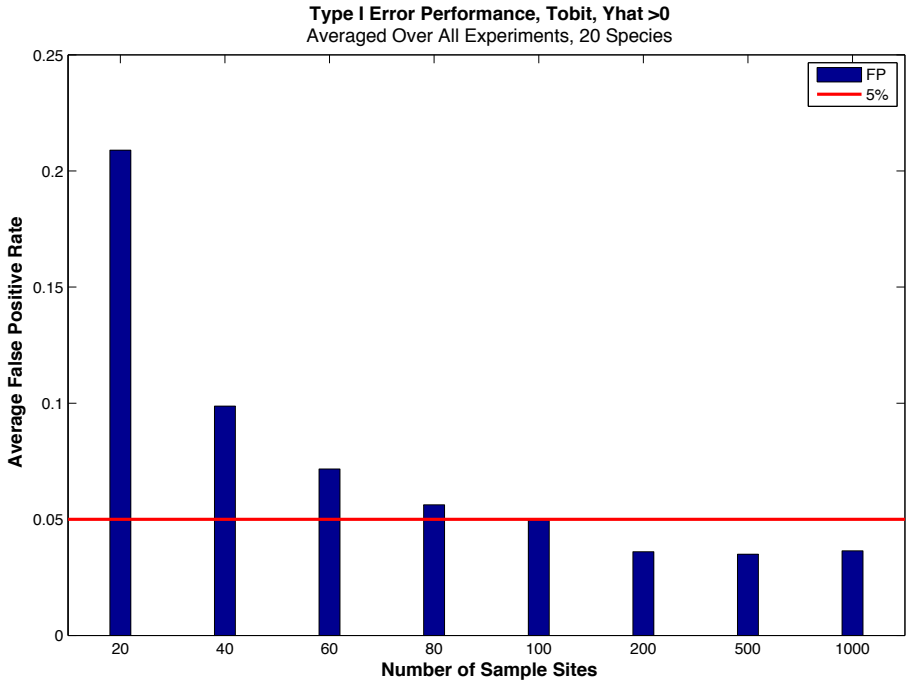

Figure 4.24: Type I error performance of Tobit regression fixed on the number of sites factors averaged over all experiments. The null model consisted of 1000 randomizations.The false positive rate is higher than SS robust regression.

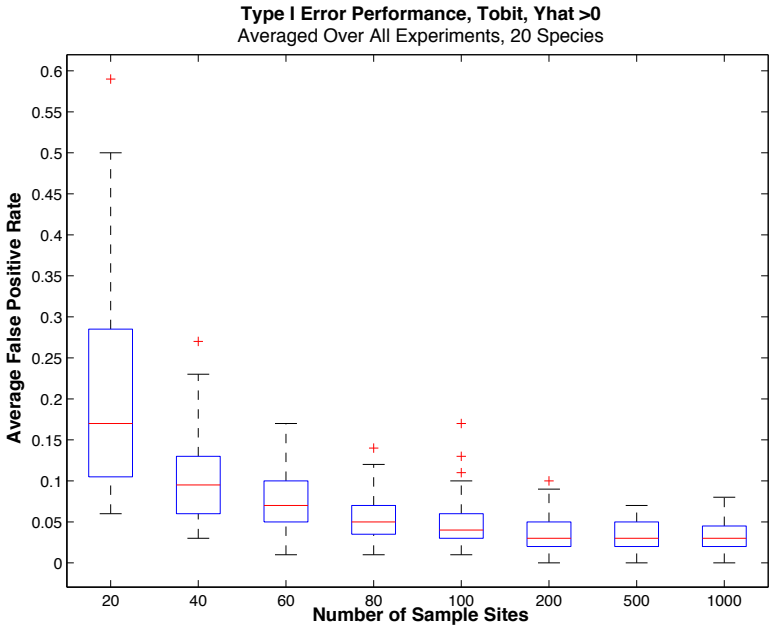

Figure 4.25: Box plot: Type I error performance of Tobit regression fixed on the number of sites. The null model consisted of 1000 randomizations. The false positive rate is higher than SS robust regression.

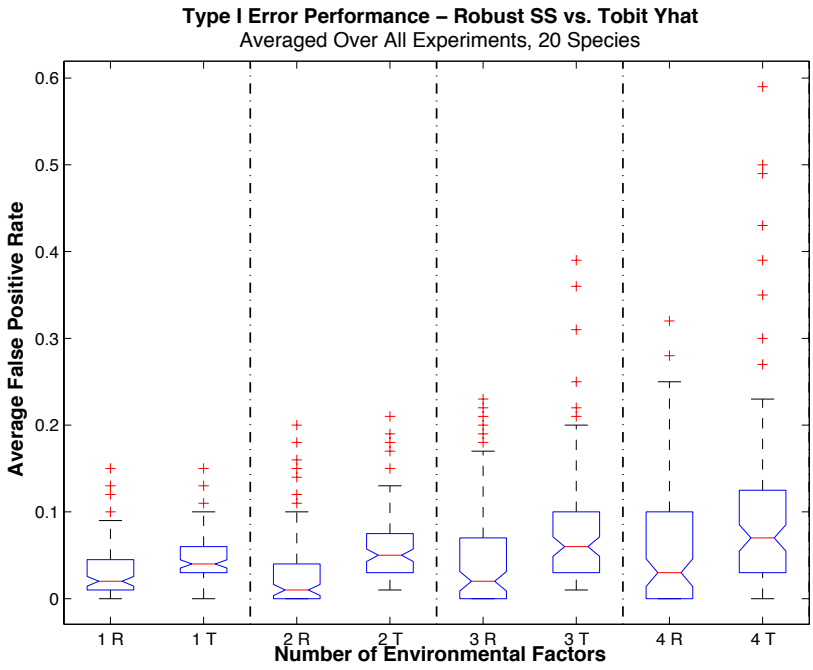

Figure 4.26: Comparison of Type I error performance between SS Robust regression and Tobit regression. The experiments are fixed on the number of environmental factors. The x-axis is labeled by the number of environmental factors and R for Robust regression or T for Tobit regression. The null model consisted of 1000 randomizations. SS Robust regression has a significantly lower false positive rate than Tobit regression.

Parameter	Values
# of Species	20
# of Env. Factors	1,2,3,4
# of Sample Sites	20, 40, 60, 80, 100, 200, 500, 1000
# of Covarying Pairs	1, 2, 10
Type of Community Covariation	Positive, Negative, Mixed
Amount of Covariation, σ_M^2	1, 3, 5, 10, 20
Noise Variance, σ_N^2	0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 2

Table 4.4: Parameter and value sweeps for the true positive detection experiments. There are a total of 17,280 parameter configurations evaluated for each experiment. Experiments repeated 100 times. The null distribution is constructed using 200 repetitions.

example, a covariation value of 1 means that σ_M^2 equals σ_N^2, a covariation of 3 means σ_M^2 equals $3\sigma_N^2$, etc. Table 4.2 summarizes the parameters and corresponding values swept in the true positive detection experiments. There are a total of 17,280 experiments conducted over the parameter sweeps. Each experiment is repeated 100 times. A community is classified as exhibiting significant species interactions if the significance value is less than 0.05.

Figures 4.27 thru 4.29 are the true positive detection counterparts to figures 4.16 thru 4.20 in the false positive analyses. These figures show the performance averaged over all parameter sweeps. The curse of dimensionality arising from the number of environmental gradients included in the study decreases the power of the test as expected. Increasing the number of sites increases power, however, and may be used to mitigate the aforementioned effect. The performance over R^2 is similar to the false positive rate in that high R^2 values are more problematic. Overall, the true positive performance appears to be acceptable over the range of environmental factors, sites and R^2 values explored.

Figures 4.30 thru 4.33 show the performance over the parameters directly affecting the community biotic interactions. Figure 4.30 shows detection performance increases with the amount of biotic interaction within the community. Figure 4.31 indicates that the magnitude of biotic covariation within the community should be greater than the amount of noise. This corresponds with a signal-to-noise (SNR) ratio greater than 1, which is acceptable since the signal should be above the noise floor. Figure 4.32 shows that increasing the number of covarying pairs within the community increases the detection rate. Figure 4.33 indicates the method has comparable performance for communities

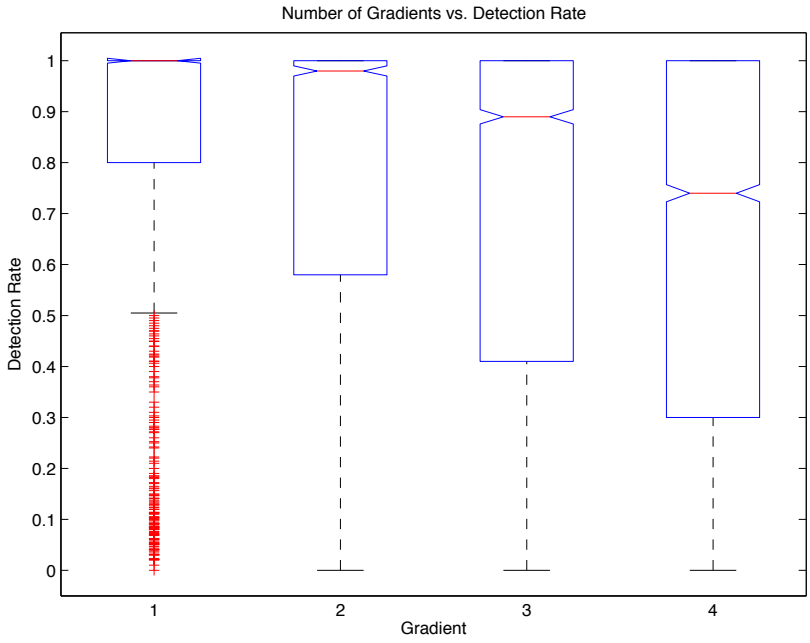

Figure 4.27: Sensitivity of SS robust fixed on number of environmental gradients included in the model.

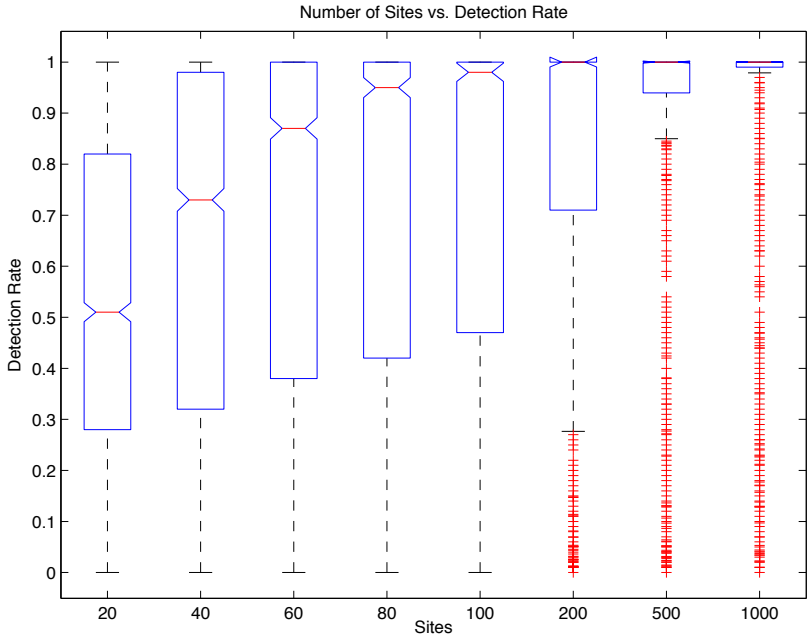

Figure 4.28: Sensitivity of SS robust fixed on number of sites included in the model.

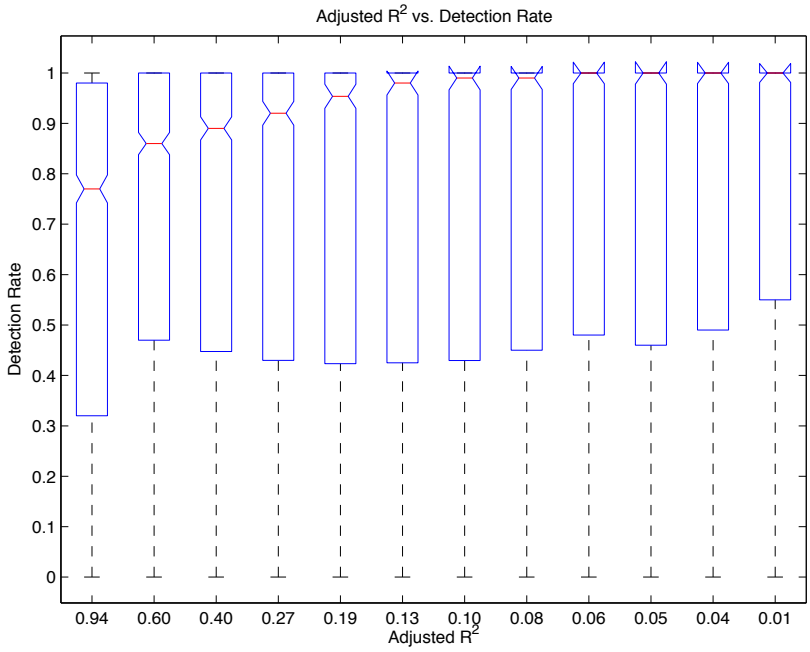

Figure 4.29: Sensitivity of SS robust fixed on the average adjusted R^2 values measured in the communities.

that are covarying strictly positively or negatively and a mixture of the two. Detection of strictly negatively covarying communities is slightly less than the others due to the effects of truncation at zero since more zeros may appear in the community matrix leading to increased uncertainty of their origin.

Figures 4.34 thru 4.36 illustrate the detection performance of positive, negative and mixed community covariation. The communities each have 20 species and one thru four environmental gradients as labeled. The y-axis refers to the SNR and the x-axis refers to the number of sites in the community matrix. The noise variance, σ_N has been fixed at 0.4 and corresponds to an R^2 of 0.19. This is indicative of values encountered in empirical datasets. Each point within the plots is superimposed with three circles. The inner, smallest circle corresponds to one covarying pair within the community, the next larger corresponds with three pairs and the largest outer circle corresponds with ten pairs. The color of the data indicates the detection rate as indicated on the color bar with

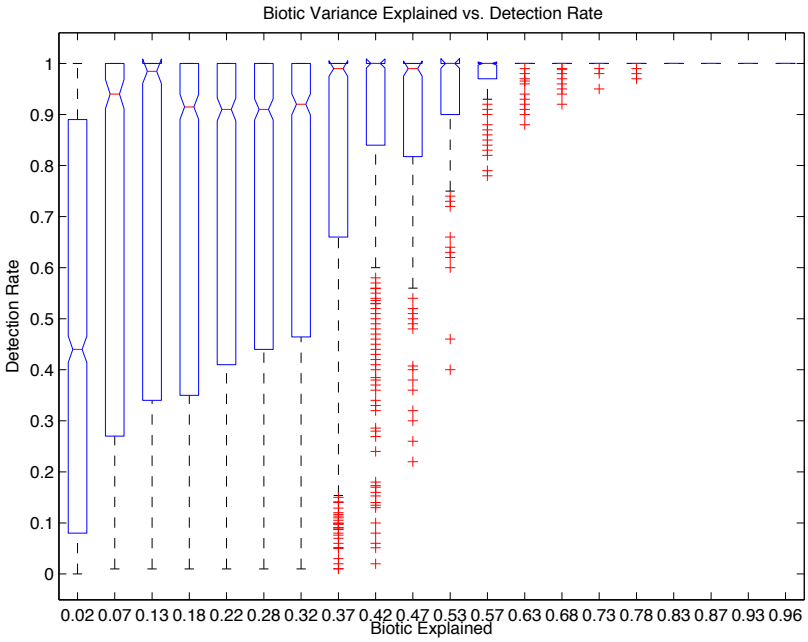

Figure 4.30: Sensitivity of SS robust display over the amount of biotic variation explained by the model.

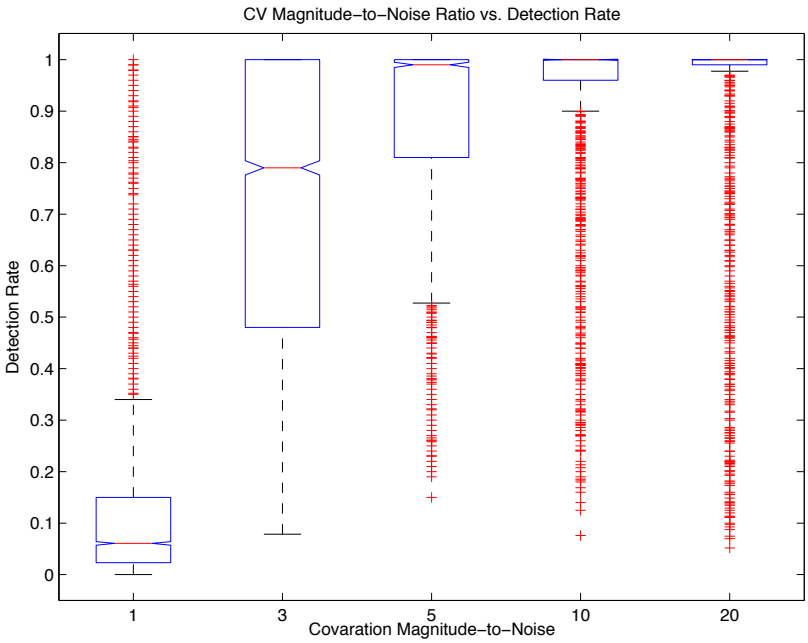

Figure 4.31: Sensitivity of SS robust with respect to the signal-to-noise ratio. The signal is the amount of covariation in the simulations as parameterized by σ_M^2.

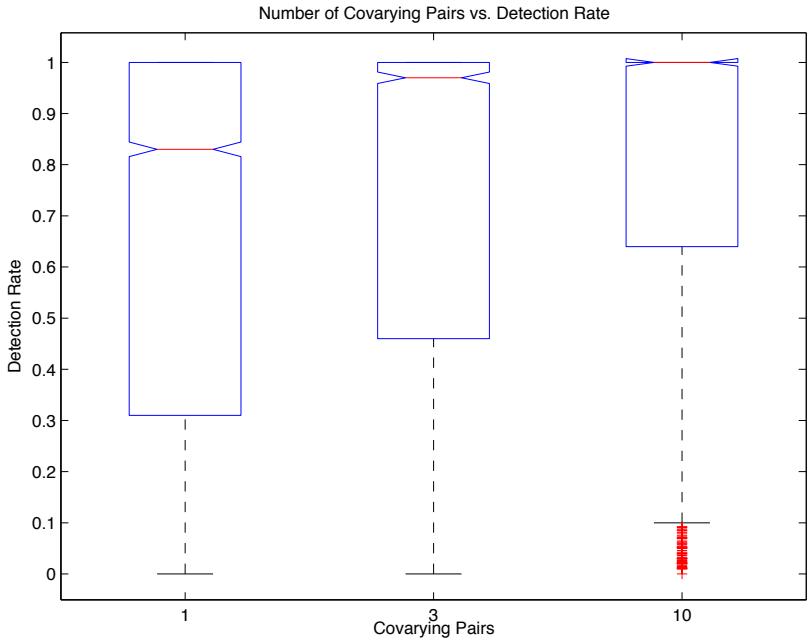

Figure 4.32: Sensitivity of SS robust fixed over the amount of covarying pairs within the community.

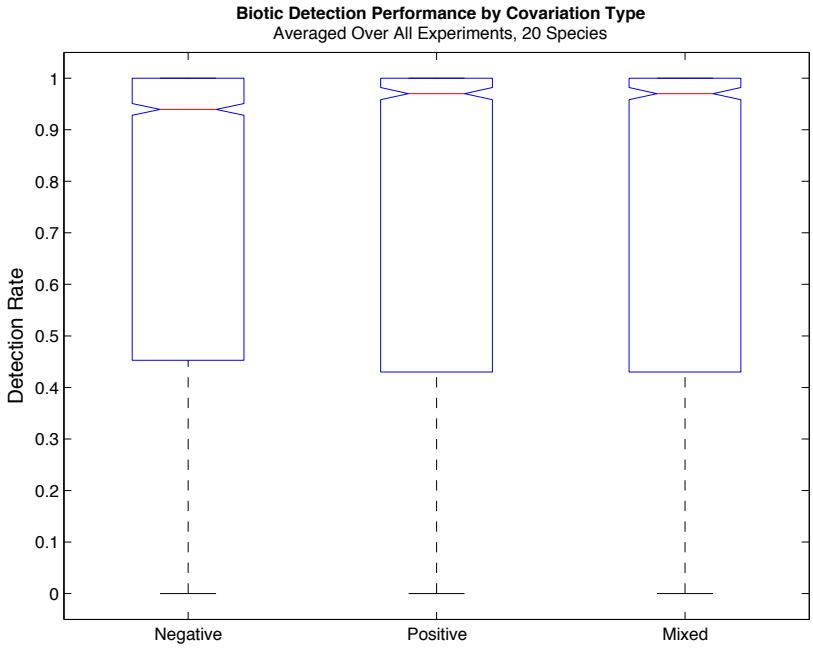

Figure 4.33: Sensitivity of SS robust fixed on type of community covariation. Mixed indicates some species pairs have positive covariation while others may have negative covariation.

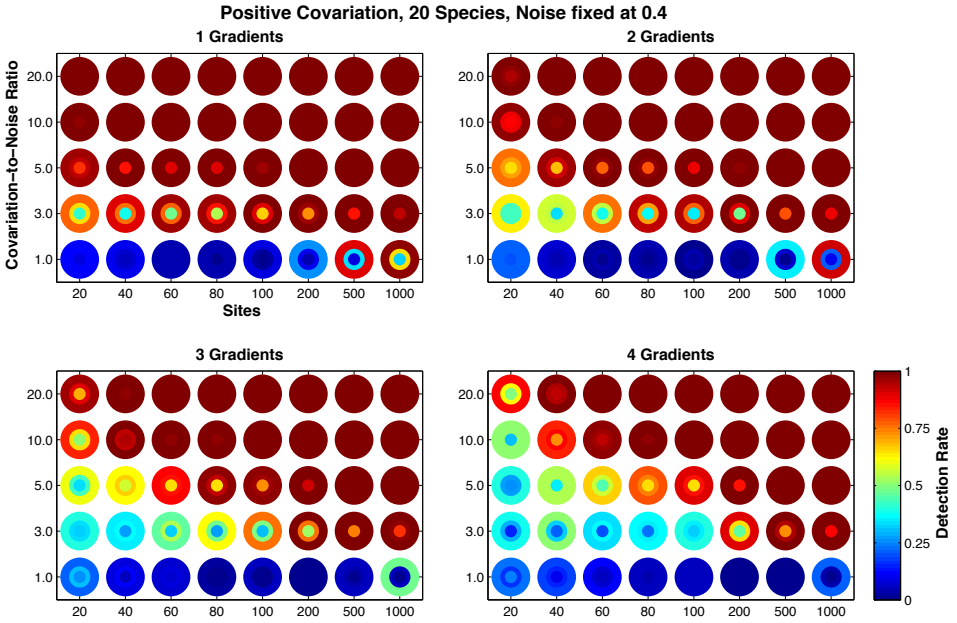

Figure 4.34: Sensitivity of SS robust regression with noise parameter σ_N^2 fixed at 0.4 and positive community covariation. The y-axis refers to the SNR and the x-axis refers to the number of sites included in the simulations. The inner most, smallest circle at each data point corresponds with 1 covarying pair in the community, the next larger 3 pairs, and the largest outer most corresponds with 10 pairs. There are a total of 20 species in the community.

the darkest red corresponding to 100% detection. A full set of similar plots illustrating sweeps over the entire noise range is provided in appendix B.

4.4 Species Parameter Sweeps

All experiments up to this point have included communities with 20 species. For many ecological studies this is a sufficient number of species to consider. However, others will include more species. Due to computational complexity and runtimes, it is infeasible to perform sensitivity sweeps on communities with a greater numbers of species in as much detail. Figures 4.37 thru 4.40 consider communities with 20, 100 and 500 species over a limited range of parameter values. This serves as a check to ensure the method is scalable and exhibits expected performance at scale.

As shown in figure 4.37 two noise values were considered providing an R^2 of 0.5 and 0.25. Sites are swept over seven values ranging from 40 to 1000 as shown. The number of environmental factors are fixed at two. Under these conditions a low false positive rate is expected. This is indeed the

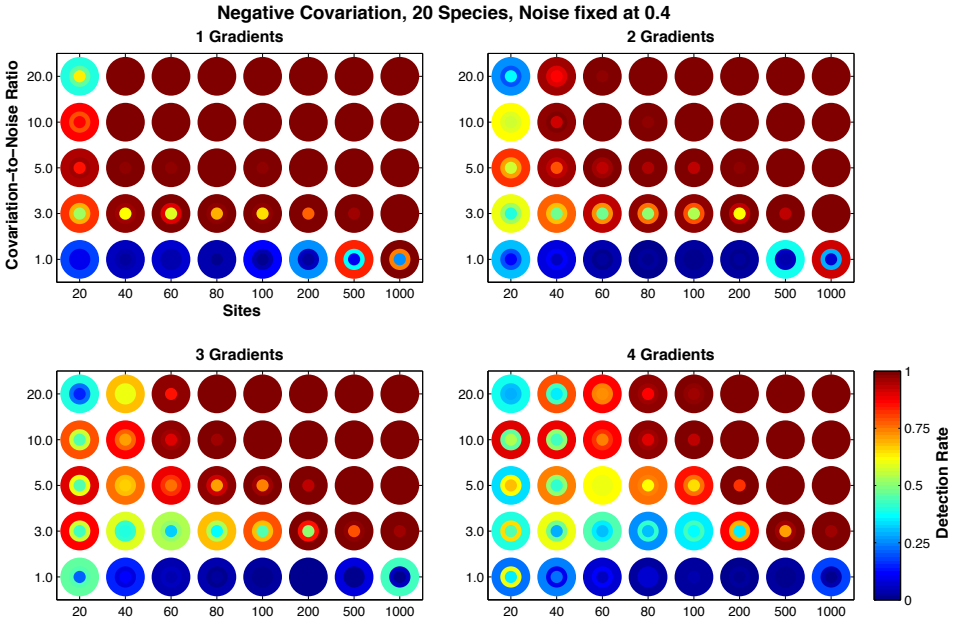

Figure 4.35: Sensitivity of SS robust regression with noise parameter σ_N^2 fixed at 0.4 and negative community covariation. The y-axis refers to the SNR and the x-axis refers to the number of sites included in the simulations. The inner most, smallest circle at each data point corresponds with 1 covarying pair in the community, the next larger 3 pairs, and the largest outer most corresponds with 10 pairs. There are a total of 20 species in the community.

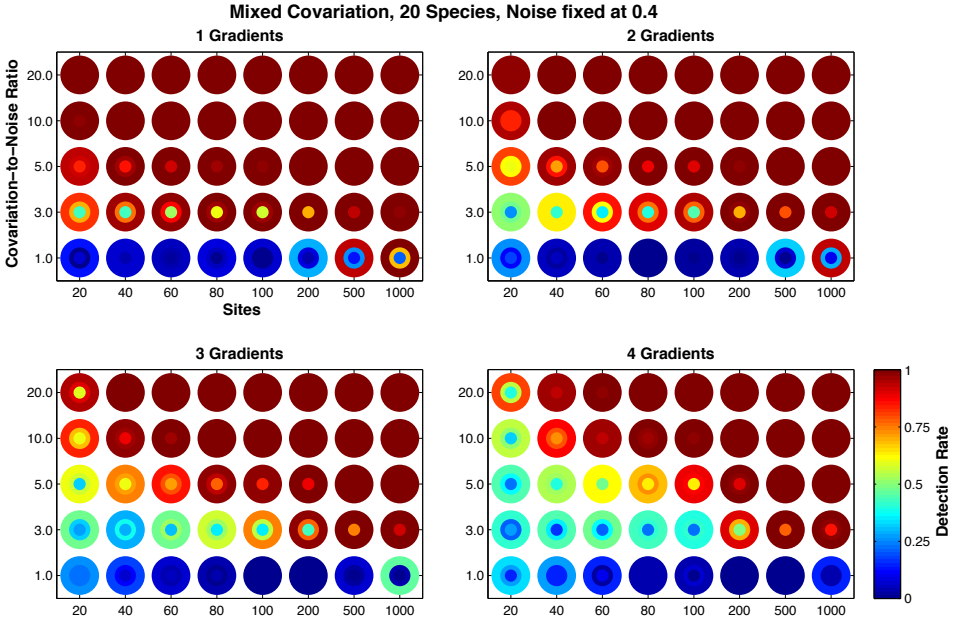

Figure 4.36: Sensitivity of SS robust regression with noise parameter σ_N^2 fixed at 0.4 and mixed community covariation. The y-axis refers to the SNR and the x-axis refers to the number of sites included in the simulations. The inner most, smallest circle at each data point corresponds with 1 covarying pair in the community, the next larger 3 pairs, and the largest outer most corresponds with 10 pairs. There are a total of 20 species in the community.

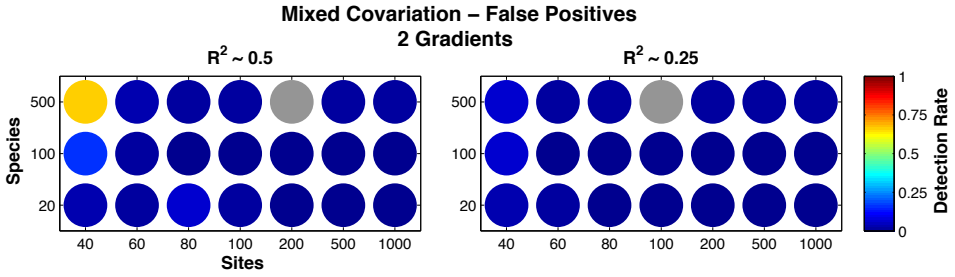

Figure 4.37: False positive rate of SS robust swept over the number of sites and number of species in the community. The experiments were fixed with two environmental factors. Two noise levels were explored as indicated by the community R^2 values. Experiments colored gray did not complete so rates are not reported.

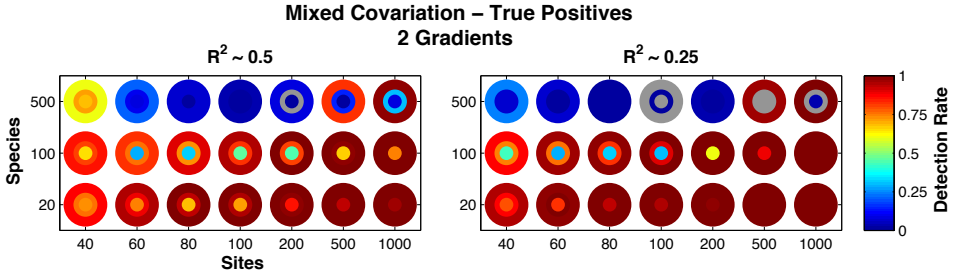

Figure 4.38: Detection rate of SS robust swept over the number of sites and number of species in the community. The experiments were fixed with two environmental factors. Two noise levels were explored as indicated by the community R^2 values. Experiments colored gray did not complete so rates are not reported.

case except for 500 species at 40 sites. This is most likely attributed to an insufficient number of sites included in the analysis with respect to the number of species. This is not an issue though as figure 4.38 shows that the ratio of sites to species should be around one for detection at the limits of detection (i.e. 10 pair out of 124,750). Increasing the ratio of covarying pairs to total pairs increases the detection rate as shown in the 100 species experiments. In these cases a site to species ratio less than one may be sufficient, but should not go below 0.4. Experiments colored gray did not have completed simulations so are thus removed from the analysis.

Figure 4.39 expands upon the 500 species experiments in figure 4.37 to include experiments swept over all 12 noise levels and increasing the number of gradients from 2 to 4. The average false positive rate is below 0.05 for experiments with more than 60 sites and quickly decreases to zero as

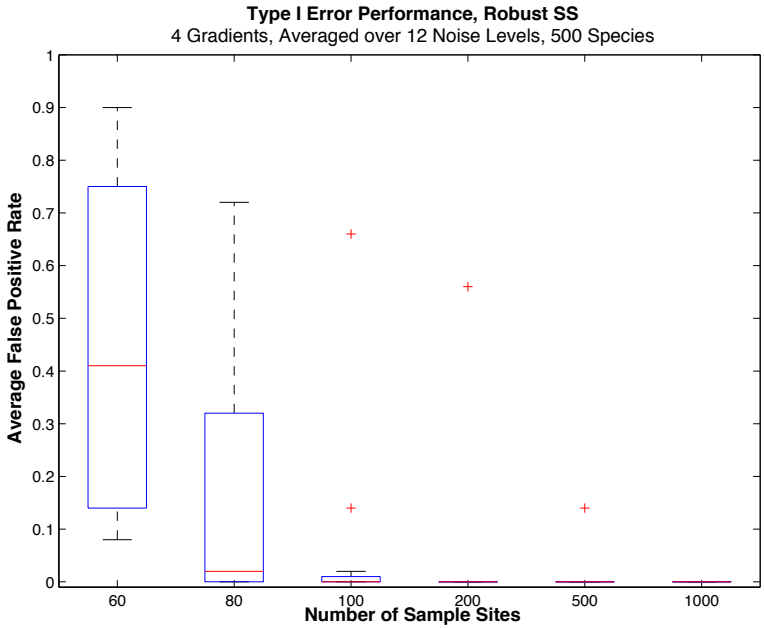

Figure 4.39: False positive rate of SS robust for 500 species experiments fixed at 4 gradients and swept over all 12 noise levels. The average false positive rate decreases to 0.05 and below as the number of sites increases beyond 60.

the number of sites increases beyond 100. Generally, a site-to-species ratio of 1 will ensure a good false positive rate in experiments with a large number of species.

Figure 4.40 is an alternative view of the 0.25 R^2 experiments in figure 4.38. This figure serves to tease out trends in detection performance. Detection rate increases with the number of covarying pairs in the community as indicated by the covary-to-total pair ratio. The greater the number of sites the better the detection rate. The more species in the community the more sites and covarying pairs are required for detection. The most important factors in achieving a high detection rate are the covarying pairs to total pairs ratio and the number of sites.

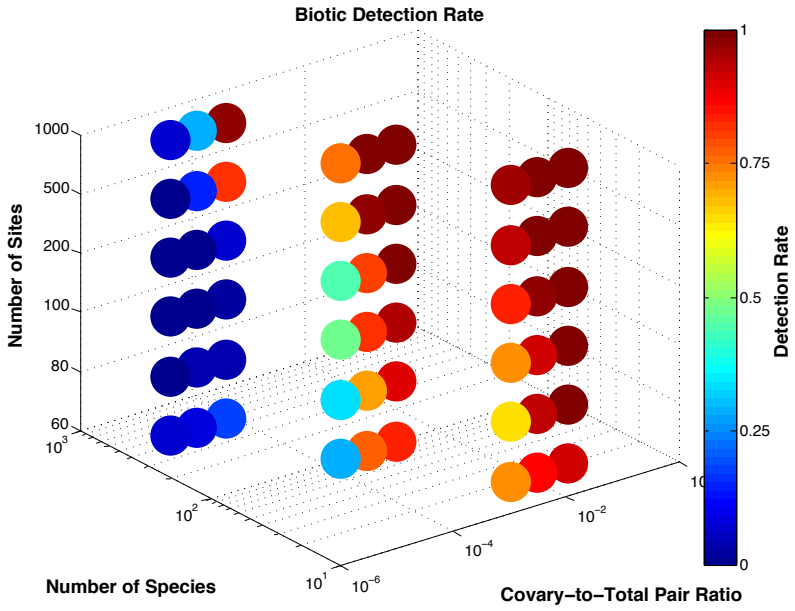

Figure 4.40: Detection rate of SS robust swept over the number of sites and number of species in the community. The experiments were fixed with two environmental factors. The noise levels was fixed with a measured community R^2 value of 0.25. The greater then number of species in the community the more covarying pairs are required for detection. The more sites included that exhibit the covariation the better the detection performance.

Chapter 5: Hybrid Method

5.1 Introduction

There is a dichotomy among methods for community structure analysis. Analyses of environmental factors influencing communities (chapter 2) and hypothesis testing of biotic interactions (chapter 3) are each carried out independently on community data. Community variation attributed to abiotic factors may be quantified as described in chapter 2, but community variation due to biotic interactions is widely unaccounted for and remains a qualitative measure based on non-parametric hypothesis testing. A hybrid approach combining the abiotic and biotic analysis has been explored in this thesis for the purposes of accounting for residual variations unexplained by abiotic factors. This permits partitioning the variation attributed to biotic interactions separately from the abiotic factors. Chapter 4 explored the performance of candidate hybrid approaches via sensitivity analyses to search for a method with suitable type I and type II error performance. This chapter summarizes the hybrid method that was selected based on its error and detection performance in the sensitivity analyses.

5.2 Hybrid Method Pipeline

The selected hybrid method shown in figure 5.1 begins with a site selection procedure to select sites that support favorable environmental conditions for each respective species. This mitigates parameter estimate biasing that frequently leads to spurious correlations among residual species responses. This procedure is followed by robust multivariate linear regression on the selected sites to estimate model parameters associated with environmental variation. A fitted response is obtained from the parameter estimates and residuals are computed from the fitted responses. The residuals are void of the environmental factors so they may be used to test for significant species interactions using a null model approach. The null model randomization and test statistic are used to generate a null distribution for which the observed residual matrix is tested for absence of significant interactions.

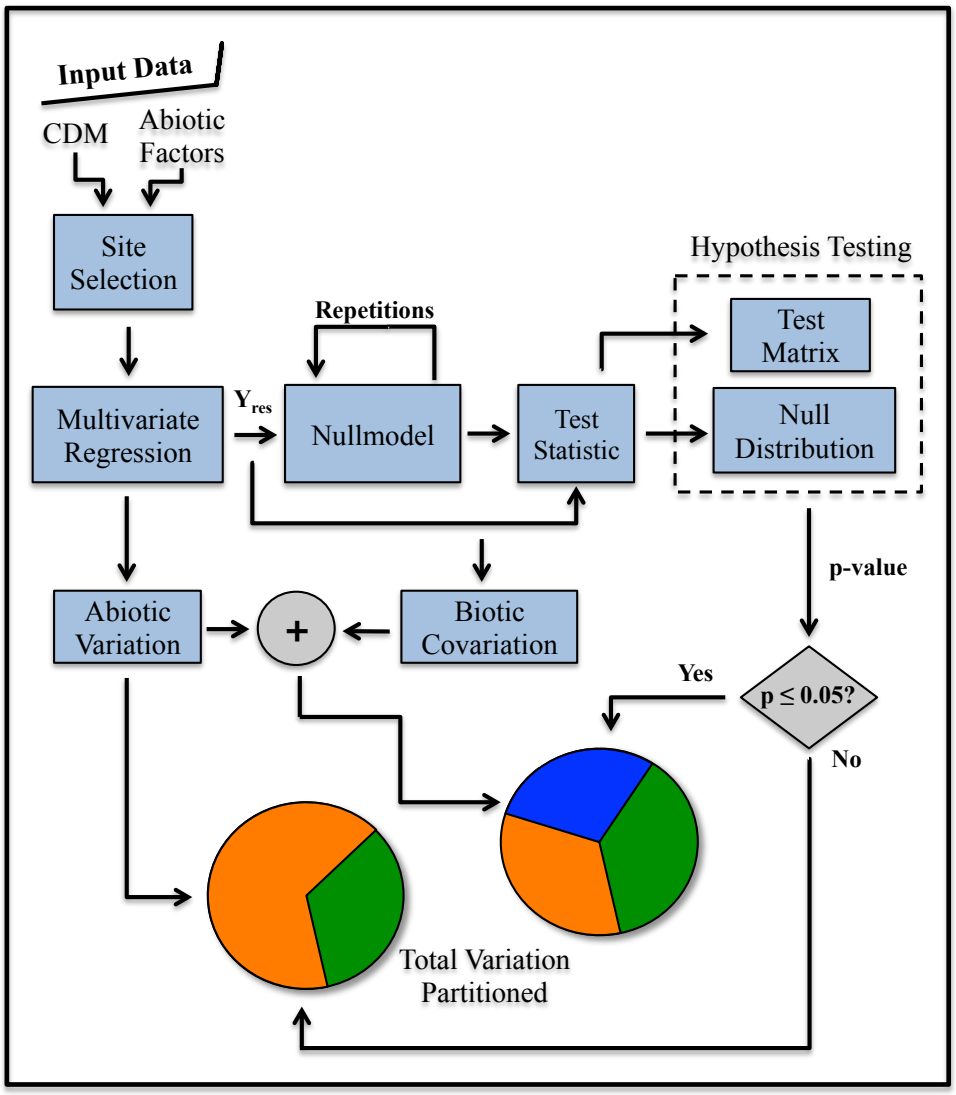

Figure 5.1: Block diagram of hybrid method combining environmental gradient analysis with species interaction hypothesis testing. A community data matrix and a set of abiotic factors are input and the variation attributed to the abiotic and biotic factors are partitioned and returned as output.

Interactions arise in the form of covariation among species pairs. These biotic interactions are quantified based on pairwise correlations if the null hypothesis is rejected. Abiotic factors are computed using the R^2 value obtained in the multivariate regression. The total variation is then partitioned and summarized. A formal description of the hybrid method is presented below.

5.2.1 Site Selection

The site selection begins by finding all sites, $\mathbf{s}$ with positive abundances for a particular species, $\mathbf{y}_k$ as in equation (5.1).

$$\boldsymbol{\zeta}_k = \{j \in \{1, 2, \ldots, n\} : \mathbf{y}_k(j) > 0\} \tag{5.1}$$

The median value of each environmental factor is computed using these sites as in equation (5.2).

$$C_i = \tilde{\mathbf{x}}_i(\boldsymbol{\zeta}_k) \tag{5.2}$$

A threshold is then computed for each factor as in equation (5.3).

$$T_i = \frac{1}{n}\sum_{j=1}^{n}(x_{ji} - C_i) + \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)}, \qquad \bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ji} \tag{5.3}$$

Sites above and below these values are omitted from further analysis as in equation (5.4).

$$\mathbf{s}_i = \{j \in \{1, 2, \ldots, n\} : (C_i - T_i) < \mathbf{x}_i(j) < (C_i + T_i)\} \tag{5.4}$$

The remaining sites that have not been discarded after analyzing each environmental factor are kept for further analysis as in equation (5.5). The procedure in equations (5.1) thru (5.5) is repeated for each of p species.

$$\mathbf{S}_k = \bigcap_{i=1}^{q} \mathbf{s}_i, \qquad \mathbf{S}_k \subseteq \{1, 2, ..., n\} \tag{5.5}$$

5.2.2 Multivariate Regression

Species response parameters, $\hat{\boldsymbol{\beta}}_k$ are estimated using the environmental factors and only those sites selected in the previous procedure as in equation (5.6).

$$\hat{\boldsymbol{\beta}}_k = (\mathbf{X}(\mathbf{S}_k, :)^T \mathbf{X}(\mathbf{S}_k, :))^{-1} \mathbf{X}(\mathbf{S}_k, :)^T \mathbf{y}_k(\mathbf{S}_k) \tag{5.6}$$

Parameter estimates, $\hat{\boldsymbol{\beta}}_k$ are used to obtain the fitted responses, $\hat{\mathbf{y}}_k$ as in equation (5.7).

$$\hat{\mathbf{y}}_k(\mathbf{S}_k) = \mathbf{X}(\mathbf{S}_k, :)\hat{\boldsymbol{\beta}}_k \tag{5.7}$$

Fitted responses with negative abundances are truncated at zero as in equation (5.8).

$$\hat{\mathbf{y}}_k(\mathbf{S}_k) < 0 = 0 \tag{5.8}$$

Residual responses are obtained from the fitted responses as in equation (5.9). The residuals are void of the environmental factors used in the regression.

$$\mathbf{y}_{\mathbf{res}_k}(\mathbf{S}_k) = \mathbf{y}_k(\mathbf{S}_k) - \hat{\mathbf{y}}_k(\mathbf{S}_k) \tag{5.9}$$

5.2.3 Null Model

The null model selected for the hybrid method randomly permutes the abundances of each species independently, with equal probability, at respective sites S_k retained in the multivariate regression site selection procedure 5.10. Once each residual species response has been permuted a test statistic is calculated for the randomized matrix. This process continues for hundreds of iterations and always starts with the original residual response matrix prior to each randomization.

$$\mathbf{y}^*_{res_k} \sim \text{random permutation of } \mathbf{y}_{res_k} \text{ over all sites} \tag{5.10}$$

5.2.4 Test Statistic

The test statistic is an index designed to consolidate the community data matrix down to a single number that summarizes the pattern of interest within the data. The index selected for the hybrid method computes the average weighted square covariation of the data matrix as in equation (5.11).

$$I_{rep} = \frac{1}{W} \sum_{k=1}^{p-1} \sum_{z=k+1}^{p} w_{zk}\, \sigma(\mathbf{y}_k^*(Sites), \mathbf{y}_z^*(Sites))^2 \tag{5.11}$$

$$w_{zk} = |\sigma(\mathbf{y}_k^*(Sites), \mathbf{y}_z^*(Sites))| \quad , \quad W = \sum_{k=1}^{p-1} \sum_{z=k+1}^{p} w_{zk} \tag{5.12}$$

For each species pair the squared covariation divided by the absolute covariation of the pair is computed using only mutual sites retained from the species site selection as in equation (5.13). The weights are the respective absolute covariation among species pairs as in equation (5.12). A threshold is set to exclude species pairs with less than seven mutual sites from the index computation. This is so covariances computed using only a few sites do not trigger detection since random covariance is more probable with a few number of sites. The average score of all species pairs is returned as the test statistic for the entire matrix.

$$Sites = \mathbf{S}_k^* \bigcap \mathbf{S}_z^* \quad , \quad rep = 1, 2, \ldots, \text{repetitions (e.g. 1000)} \tag{5.13}$$

5.2.5 Significance Test

To check if covariation observed among residual species pairs is more extreme than expected by chance the distribution of test statistics for the randomized matrices are compared to the test statistic calculated for the residual response matrix. This generates a p-value for the hypothesis test as in equation (5.14).

$$p_{value} = \frac{\#\ of\ I_{rep} \geq I_{test}}{\#\ \text{of repetitions (e.g. 1000)}} \tag{5.14}$$

A 5% significance level has been adopted for this method. For example, if there are 999 randomized permutations in the null distribution there must be no more than 50 indices that score higher than the test statistic of the observed matrix to reject the null hypothesis.

$$\begin{cases} H_1 & \text{if } p_{value} < 0.05 \\ H_0 & \text{otherwise} \end{cases} \tag{5.15}$$

5.2.6 Variation Partitioning

The explained community environmental and biotic variations are partitioned in a hierarchical manner. The abiotic variation is computed first using the coefficient of determination, R^2 as in equations (5.16) thru (5.19). The biotic variation is then computed as in equation (5.20). The final partitioning of variation is described by equation (5.22).

Abiotic Variation

Abiotic variation is computed using the R^2 value obtained from the multivariate regression as described in equations (5.16) thru (5.19). The R_k^2 values are averaged over all species and weighted by each species total abundance. The average R^2 is then adjusted to correct for multiple factors in the regression.

$$SS_{tot_k} = \sum_{j=1}^{n}(y_{jk} - \bar{y}_k)^2 \quad , \quad SS_{res_k} = \sum_{j=1}^{n}(y_{jk} - \hat{y}_{jk})^2 \quad , \quad \bar{y}_k = \frac{1}{n}\sum_{j=1}^{n} y_{jk} \tag{5.16}$$

$$R^2_{avg} = \frac{1}{W}\sum_{k=1}^{p} w_k R_k^2 \quad , \quad R_k^2 = 1 - \frac{SS_{res_k}}{SS_{tot_k}} \tag{5.17}$$

$$W = \sum_{k=1}^{p} w_k \quad , \quad w_k = \sum_{j=1}^{n} y_{jk} \tag{5.18}$$

$$R^2_{adj} = R^2_{avg} - (1 - R^2_{avg})\frac{q}{n - q - 1} \tag{5.19}$$

Biotic Variation

Biotic covariation of residual responses may be quantified based on the outcome of the null model significance test using a pairwise covariation model. Community rather than individual variation is partitioned due to vanishingly small significance thresholds (i.e. p-value) arising from corrections for multiple comparisons (e.g. Bonferroni where $p_{value} = \frac{\alpha}{p(p-1)/2}$) [41]. The necessary correction would drive the required p-value for detection towards zero causing the method to loose substantial detection sensitivity.

If the null hypothesis is not rejected then the biotic variation is zero so the remaining variance is returned as unexplained variation (i.e. Unexplained = 1 - R^2_{adj}). If the null hypothesis is rejected, however, then variation due to biotic interaction is reported. This variation is obtained by averaging the maximum squared correlations among all valid species pairs as in equation (5.20). This is then multiplied by the variation unexplained by the environmental factors as in equation (5.22).

$$\mathrm{B} = \frac{1}{W}\sum_{k=1}^{p} \max_{z} \rho(\mathbf{y}_k(\mathbf{S}_k), \mathbf{y}_z(\mathbf{S}_z))^2 w_k, \quad \text{Biotic} = (1 - R^2_{adj})B \tag{5.20}$$

$$W = \sum_{k=1}^{p} w_k, \quad w_k = \sum_{j \in \mathbf{S}_k \bigcap \mathbf{S}_z} y_{jk} \tag{5.21}$$

$$R^2_{adj} + \text{Biotic} + \text{Unexplained} = 1 \tag{5.22}$$

5.3 Preprocessing

The hybrid method does not require much preprocessing, but a there are a few points to consider when attempting to performance an analysis. Centering of the dependent (e.g. species responses) and/or the independent (e.g. environmental) variables is unnecessary since an intercept parameter is estimated. Species responding linearly to environmental conditions (i.e. short gradients) are appropriate for analysis, while those with unimodal responses (i.e. long gradients) are inherently non-linear and should be avoided.

5.3.1 Absolute Abundance

Only raw abundance species responses should be included in the biotic covariation analysis. Detection of covariation among relative abundance responses is essentially a misnomer. This is due to the inherent correlation introduced by the sampling and normalization process [42].

5.3.2 Data Transformations

Distances among objects produced by RDA are inherently euclidean. There has been much discussion in the literature, however, about the insufficiency of the euclidean distance for analyses involving species count data with null abundances [43, 17, 11]. Issues such as double-zero symmetry, species-replication variance and failure to produce the largest dissimilarity for pairs of sites with non-overlapping species have led researchers to develop approaches to mitigate the adverse effects produced by the euclidean distance present in RDA based analyses [13, 17, 11]. For example, distance-based RDA was developed to allow any suitable distance (metric or non-metric) for species data to be used for community composition analysis at the expense of producing non-linear combinations of species responses. This non-linearity obscures the contribution of each species to the ordination and is therefore not a viable option for interspecific interaction analysis.

The use of data transformations prior to RDA have been broadly advocated to diminish distortion arising from the use of the euclidean distance [43, 17, 11]. While a consensus is difficult to obtain due to diverging analytical goals among researchers, the central theme is to transform data prior to ordination or beta diversity analysis. For example, the Hellinger distance has been recommended as a distance measure for the clustering and ordination of species abundance data due to superior distance preservation among objects when moving from high to low-dimensional space [44]. Correspondingly, the Hellinger transform (square root of relative abundance) may be applied to data prior to ordination. The resulting data may then be subjected to RDA analysis since the euclidean distance of the transformed data is equivalent to the Hellinger distance of the raw data, thus circumventing the issues encountered with the euclidean distance in RDA [43].

A simulation study of the Hellinger distance showed that the transformation applied to data prior to RDA improved the estimate of the multiple coefficient of determination on normally distributed

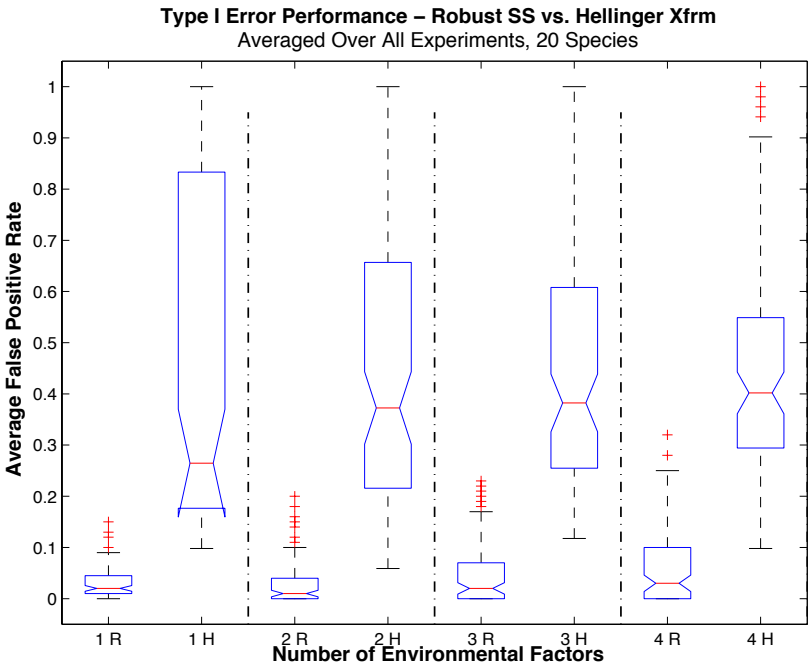

Figure 5.2: Box plot: Type I error performance of SS robust regression fixed on the number of environmental factors. The x-axis is labeled by the number of environmental factors and R for robust regression or H for preprocessing with the Hellinger transform. The null model consisted of 1000 randomizations. This figure shows that the transformation significantly increases the false positive rate and should not be used with the present method.

data [17]. While this outcome is most useful for ordination and beta diversity purposes, it should be noted that this transformation is not appropriate prior to using the present detection method. The Hellinger transformation inherently introduces correlation among species responses at each site. This leads to unsuitable false positive rates of detection as can be seen in figure 5.2. In general, non-linear transformations should not be used prior to analysis due to the distortion imposed by the transformation on response residuals.

5.3.3 Variable Selection

An explanatory variable selection procedure should be employed so that relevant environmental variables are included in the regression [45]. This will assist in removing collinearity among these variables. Correlations among species residual responses will be detected if relevant explanatory variables of large effect are omitted from the regression.

5.4 Computational Complexity Benchmark

An analysis of the computational complexity of the algorithm is useful for understanding how the method will scale with large communities. For empirical data the complexity is not as much of an issue as it is for sensitivity analyses. There are over 17,000 experiments needed to perform a single sweep over the parameters in the sensitivity analysis in chapter 4. Each experiment is comprised of 100 iterations each generating a different community matrix. Thus there are a total of 1.7 million community matrices required to complete the sensitivity analysis.

Each iteration requires at least 200 randomization repetitions to generate the null distribution based on the discussion in chapter 3. The test statistic computation is the bottleneck for the analysis pipeline since it is computed among all species pairs in the community. The computation complexity average and worst-case performance scales with $O(p^2)$ where p is the number of species in the community. A benchmark of the wall-clock performance is provided in figure 5.3. This benchmark was completed using a linux server running OpenSuse 12.2 with 548 GB of RAM and 152 available cores from Intel Xeon E7 4820 2.0 GHz processors.

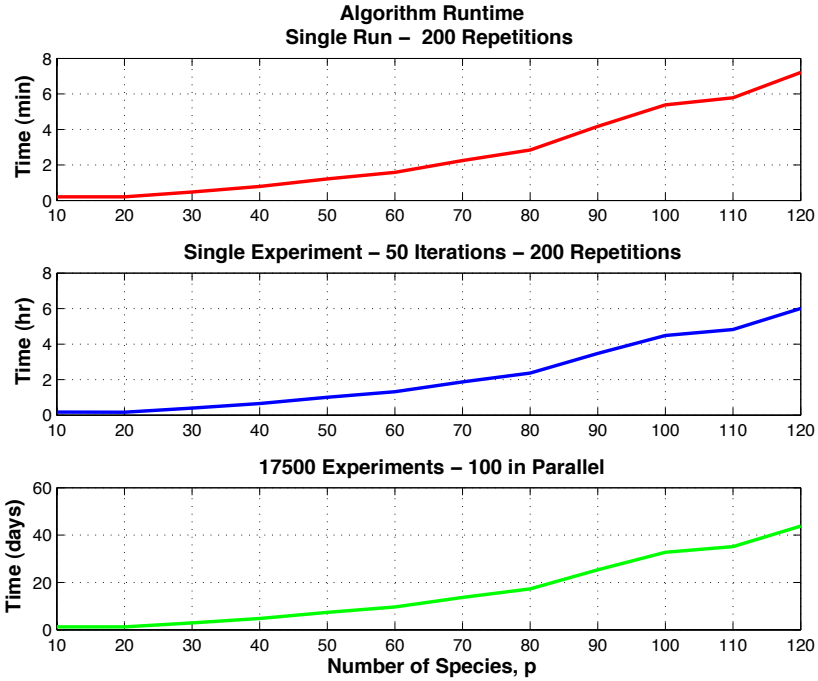

Figure 5.3: Benchmarking the wall-time computational performance of the hybrid method. The top plot refers to a single run with 200 repetitions for creating the null model. The middle plot shows the average run-time of an experiment consisting of 50 iteration. The bottom plots shows the performance for 17,500 experiments as needed for the sensitivity analysis. Note that this assumes there that 100 experiments run in parallel on 100 CPU cores.

Chapter 6: Case Study - Leaf Litter Diversity

6.1 Introduction

Analyses and discussions over the previous four chapters have been driven by the need for a hybrid approach that combines environmental factor analysis and biotic interaction hypothesis testing for the purposes of quantifying and elucidating the influences behind observed ecological community structure. A suitable hybrid method has been developed based on sensitivity analyses of a set candidate solutions as described in chapter 4. The details of the selected method as implemented is presented in chapter 5. Now that a qualified method is available a case study illustrating an example of a typical application of its use in practice is now presented for the first time.

This chapter begins with an overview of a published leaf litter diversity study, the significance of the research and applicability of the hybrid method's use on the study. The suitability of the data collected is discussed. Analysis using the method is shown followed by a brief discussion interpreting the results and providing suggestions for future qualifying and exploratory experiments.

6.2 Overview

Populations of species form a community of organisms and multiple communities form an ecosystem [10]. Questions such as those surrounding the impacts of climate change at the ecosystem level ultimately depend upon the constituent organisms of the communities and their respective populations. The influence of the environment and species interactions on community structure is therefore a central topic in ecology.

Theories about community structure range from the Clemantsian superorganismic view where species interactions are assumed to play a major role in development and regulation [46] to Gleason's individualistic concept where species interactions have a minimal role in community structure [47]. Factors behind empirical community structure lie somewhere on a spectrum between these two extreme views with the effects on structure differing among communities. The impact of the envi-

ronment on community structure is well-known so the effect size of which specific factors influence community structure is the focus of many studies. The next step is to remove the environmental factors that explain variations in community structure and then determine what portion of remaining community variation is attributed to biotic interactions. The hybrid method enables this next step in community structure analysis and the study described below serves as motivation.

6.2.1 Motivation

Saprotrophic microorganisms break down dead organic matter and produce carbon (C) and nitrogen (N) byproducts that serve as nutrients for the ecosystem [10]. For example, in a forest, carbon and nitrogen are introduced into the soil by microorganisms consuming dead leaf matter. The C and N are then consumed by the roots of a tree to fuel cellular processes for the tree to function. The tree grows and produces leaves, which fall back to the soil when they die and the cycle continues. The tree produces oxygen and other byproducts that also contribute to the ecosystem. Given that this is a delicate, well-balanced system, understanding the consequences of shifts in biodiversity is important for predicting the impacts of perturbations to factors for ecosystem stability.

6.2.2 Synopsis

The case study described herein begins with a published study on the effects of leaf litter diversity on saprotrophic microorganism communities [48]. Leaf litter diversity refers to the quantity and type of leaves in the soil communities. Leaves are microcosms with bottom-up effects to the rest of the ecosystem as described above. Prior to this study it was unclear whether litter diversity had a predictable impact on the soil microbial communities. The study controlled for leaf type over two forest habitats. Several relevant environmental factors thought responsible for structuring the communities were also measured.

Community profiles were obtained from litter-bags containing either one species or four different species of leaves. This is a controlled experiment so other factors were fixed. RDA was used to asses the amount of variation observed within the microbial community as explained by the leaf type and habitat. Significant variability within the community was explained by leaf types. It was also found

that the diversity of leaves did not explain significant variability thus it is the type of leaves present that contribute to microbial community structure. This is an important finding since leaf type must be considered in other litter diversity studies so that findings of diversity effects among communities is not confused with other factors in the studies.

The results of this study showed that a large portion of variation (approximately 50%) remains unexplained by the environmental factors, including leaf type. It is hypothesized that biotic interactions may explain a portion of the currently unexplained variation. This hypothesis may now be addressed using the hybrid method.

6.3 Hybrid Method Analysis

The community data matrix for the study consists of 87 species sampled at 264 different sites. It is desirable to have at least as many sites as there are species in the analysis based on the findings of the sensitivity analysis in chapter 4. This data matrix exclusively contains bacterial species and their absolute abundance counts as described below.

6.3.1 Suitability of Data Collected

Microbial community profiles were obtained using TRFLP [49]. Each microbe has a 16S rRNA gene that may by used to discriminate among different species based on the nucleotide sequence [50]. TRFLP is a fingerprinting method that cuts the sequence at a restriction site and then amplifies the sequences by polymerase chain reaction PCR [51]. The sequences are then sequenced and characterized based on their length. TFRLP can be used to infer the diversity of a microbial community and the relative abundance of each species. However, additional information is required to compute the relative abundance to absolute abundance. This was accomplished by obtaining the microbial biomass at each sample site and then multiplying the biomass times the TRFLP results.

The choice of environmental factors is an important consideration in community analysis and must be selected based on knowledge of the particular domain. Omitting factors having large contributions to community structure may produce misleading results. Furthermore, those factors will be apparent in the residuals after regression and therefore trigger detection of significant covariation.

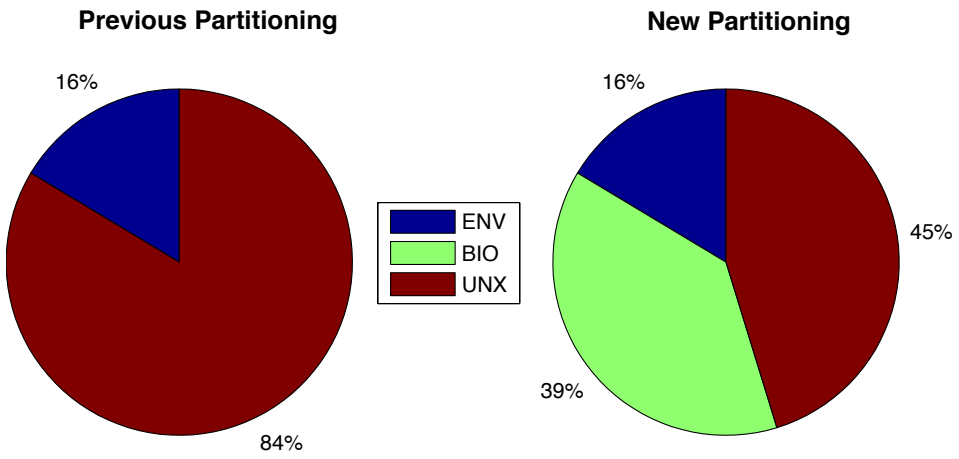

Figure 6.1: Variance partitioning of community averaged data as returned by the hybrid method. The community has 264 sites, 87 species and 4 environmental factors. There were 1000 randomization used to create the null model distribution. The p-value is 0.002 indicating significant covariation among community response residuals. Note that the hybrid method is able to account for 46% of the previously unexplained variation (i.e. 39/84).

An environmental variable selection procedure was used to select variables contributing to community variation [45]. This procedure also omits others that are not significant and may exhibit collinearity with the selected variables. Four environmental variables were selected based on this procedure: C/N Ratio, proportion_pol compounds, cellulose, and lignan content.

6.3.2 Results

Figures 6.1 and 6.2 present the results of the analysis. Figure 6.1 shows the community averaged data, while figure shows the variance explained for each species response. There were 1000 randomizations used to create the null distribution for the hypothesis test. Significant covariation among the response residuals was found at p-value of 0.002. The environmental explained variation is approximately 13%. A large amount of biotic interaction was found accounting for 38% of the total variation.

Note that the figure 6.2 does not show significant interaction at the p-value found in the hypothesis test since the p-value is computed for the entire community rather than individual responses as described in chapter 3 and 5.

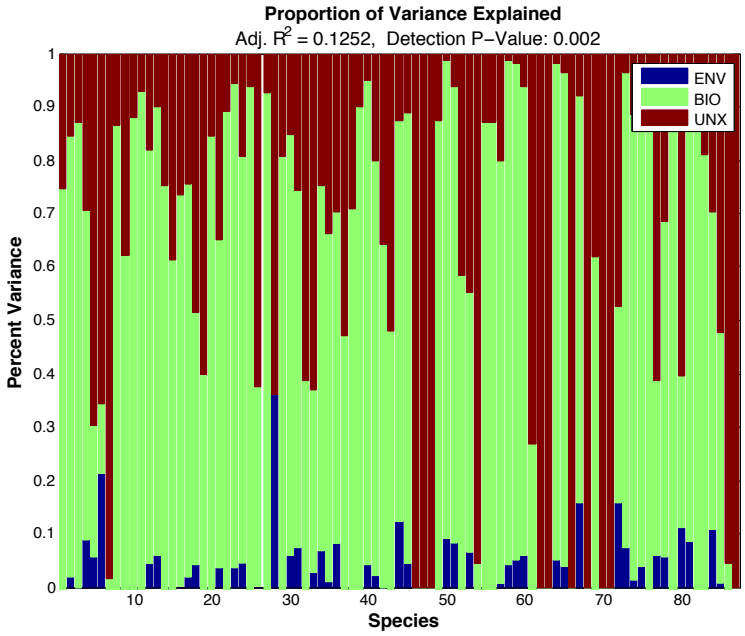

Figure 6.2: Variance partitioning of individual species as returned by the hybrid method. The variation of the 87 species shown here is for hypothesis generating purposes. The hypothesis test returning the p-value of 0.002 is only valid for the community averaged data.

6.3.3 Discussion

The environmental factors were shown to account for about 13% of total variation. Note that this value is lower than the RDA R^2, which is approximately 22%. The discrepancy arises since the RDA includes all sites in the regression and is arguably a biased estimate. Significant interaction among species pairs in the community was found as suspected based on the results of the original study. This is an exciting result because biotic interactions are accounting for a large portion of the previously unexplained variation. It remains to be found the cause of the biotic interactions as they may arise from competitive interaction, mutualism, ecosystem engineering, etc. The results of this analysis should serve as motivation for subsequent experimental studies to target specific organisms and probe for origins of smaller scale variations.

Chapter 7: Conclusion

A rigorous method for testing biotic interactions using species count data void of responses to environmental factors has been developed. Output from the method provides partitioning of variances attributed to abiotic and biotic factors. The results may be used for exploratory data analysis and may be use to drive further experimental work. The method is publicly available for download as an R package [http://github.com/EESI/NullSens].

A number of challenges were encountered in the development of the hybrid method. Regression on truncated data such as species counts is problematic in biasing the regression parameter estimates and for introducing correlations between the predicted and residual response. This ultimately leads to correlations among the residuals and contributes to high false positives in subsequent biotic interaction hypothesis testing. Alternative regression methods were explored and a site selection procedure was developed that draws upon domain knowledge about the origin of zeros in community data. Non-parametric hypothesis testing was introduced via a null model developed for detecting significant covariation among species residual responses. This formulation of the accompanying test statistic was designed around a pairwise species interaction model and circumvents the issue of correcting for multiple comparisons. The full description of the hybrid method was presented in chapter 5.

The sensitivity analysis of the chosen method showed an acceptable tradeoff of type I and type II errors over a wide range of parameters typically encountered on real datasets. These parameters include number of sites, species and gradients (environmental factors), noise, number covarying species pairs, the amount of covariation among the pairs and the type of covariation (positive, negative or mixed).

Based on acceptable performance in the sensitivity analysis the method was used in a case study on leaf litter diversity. The previously published method hypothesized that much of the residual variation unexplained by the environmental facts was attributed to species interactions. The case

study shown in this thesis found that this was true and may serve as motivation for subsequent leaf litter analyses.

7.0.4 Future Research

The method as presented provides a framework for abiotic and biotic analyses. Future research should be directed toward incorporating non-linear species responses into the regression model. The capability to handle relative abundance data would be very beneficial, especially in studies involving microbial communities where biomass has not been measured. The investigation of hierarchal species interactions should also be explored for use in datasets where it may be a more appropriate reflection of the data than the pairwise interaction model.

Finally, the analysis of latent covariation masked my explanatory variables need not be restricted to applications in community ecology. The methods developed within this thesis are also applicable to other domains where censored data has been observed. Future investigations should search for additional applications that may benefit from the developments in this thesis. To start, the methods may most readily find application in finance and economics.

Appendix A: Mathematical Notation

Parameter	Description	Size
$a^{(c)}$	Species 1 in c^{th} Pair of Interacting Species	1 x 1
$b^{(c)}$	Species 2 in c^{th} Pair of Interacting Species	1 x 1
b_{ki}	Species Response Parameter, k^{th} Species to i^{th} Gradient	1 x 1
$\mathbf{B}$	Species Response Parameter Matrix	(q+1) x p
$\hat{\mathbf{B}}$	Species Response Parameter Matrix Estimate	(q+1) x p
Biotic	Biotic Explained Variation	1 x 1
C_i	Centroid of $\mathbf{x}_i$	1 x 1
I_{rep}	Test Statistic Value for Repetition of Null Model	1 x 1
$\mathbf{M}$	Species Interaction Matrix	n x p
n	Number of Sites	1 x 1
$\mathbf{N}$	Noise Matrix with N_{jk} Noise for k^{th} Species at j^{th} Site	n x p
p	Number of Species	1 x 1
pool	Non-Interacting Species	1 x variable
q	Number of Gradients (Environmental Factors)	1 x 1
R_k^2	k^{th} Species Coefficient of Determination (CoD)	1 x 1
R_{adj}^2	CoD Adjusted for Multiple Gradients	1 x 1
R_{avg}^2	Community Averaged CoD, Environmentally Explained Variation	1 x 1
$\mathbf{s}_i$	Sites for Inclusion for i^{th} Gradient	1 x variable
$\mathbf{S}_k$	Sites for Inclusion for k^{th} Species	1 x variable
Sites	Sites selected for Test Statistic Calculation	1 x variable
T_{min_i}	Threshold: Minimum Value of i^{th} Gradient for Inclusion	1 x 1
T_{max_i}	Threshold: Maximum Value of i^{th} Gradient for Inclusion	1 x 1
v	Number of Interacting Species Pairs	1 x 1
w_k	Total Abundance of k^{th} Species	1 x 1
w_{zk}	Weight for Species Pair z,k	1 x 1
W	Sum of Weights	1 x 1
x_{ji}	Value of i^{th} Gradient at j^{th} Site	1 x 1
$\bar{x}_i$	Mean of i^{th} Gradient	1 x 1
$\tilde{x}$	Median of i^{th} Gradient	1 x 1
$\mathbf{X}$	Environmental Gradient/Factor Matrix	n x (q+1)
$\mathbf{X}(\mathbf{S}_k, :)$	Gradient Matrix only at Selected Sites for k^{th} Species	variable x q
y_{jk}	k^{th} Species Abundance at j^{th} Site	1 x 1
$\bar{y}_k$	k^{th} Species Average Abundance	1 x 1
$\hat{\mathbf{y}}_k$	k^{th} Species Predicted (Fitted) Response	n x 1
$\hat{\mathbf{y}}_k(\mathbf{S}_k)$	Predicted Abundance for k^{th} Species at Selected Sites	variable x 1
$\mathbf{y}_{res_k}$	k^{th} Species Residual Response	n x 1
$\mathbf{y}^*_{res_k}$	Random permutation of $\mathbf{y}_{res_k}$	n x 1
$\mathbf{y}_{res_k}(\mathbf{S}_k)$	Residual Abundance for k^{th} Species at Selected Sites	variable x 1
$\mathbf{Y}$	Community Data Matrix	n x p
$\hat{\mathbf{Y}}$	Predicted (Fitted) Response Matrix	n x p
$\mathbf{Y}_{res}$	Residual Response Matrix	n x p
$z_j^{(c)}$	Interaction Abundance for c^{th} Species Pair at j^{th} Site	1 x 1
ρ	Correlation	1 x 1
σ	Covariance	1 x 1
$\boldsymbol{\zeta}_k$	Set of Sites with Positive Abundance for k^{th} Species	1 x variable

Appendix B: Mixed Community Detection Results

Figures B.1 thru B.12 below are an expansion of the results presented in figure 4.36 in the sensitivity analysis presented in chapter 4. Each figure illustrates the detection performance of communities with mixed (positive and negative) covariation. The communities each have 20 species and one thru four environmental gradients as labeled. The y-axis refers to the SNR (covariation magnitude to noise magnitude) and the x-axis refers to the number of sites in the community matrix. The noise variance, σ_N has been fixed within each plot. These values have been chosen to be indicative of values encountered in empirical datasets as summarized in table 4.3. Each point within the plots is superimposed with three circles. The inner, smallest circle corresponds to one covarying pair within the community, the next larger corresponds with three pairs and the largest outer circle corresponds with ten pairs. The color of the data indicates the detection rate as indicated on the color bar with the darkest red corresponding to 100% detection. Each experiment is the average of 100 iterations. There were 200 randomizations used to generate the null model in each iteration.

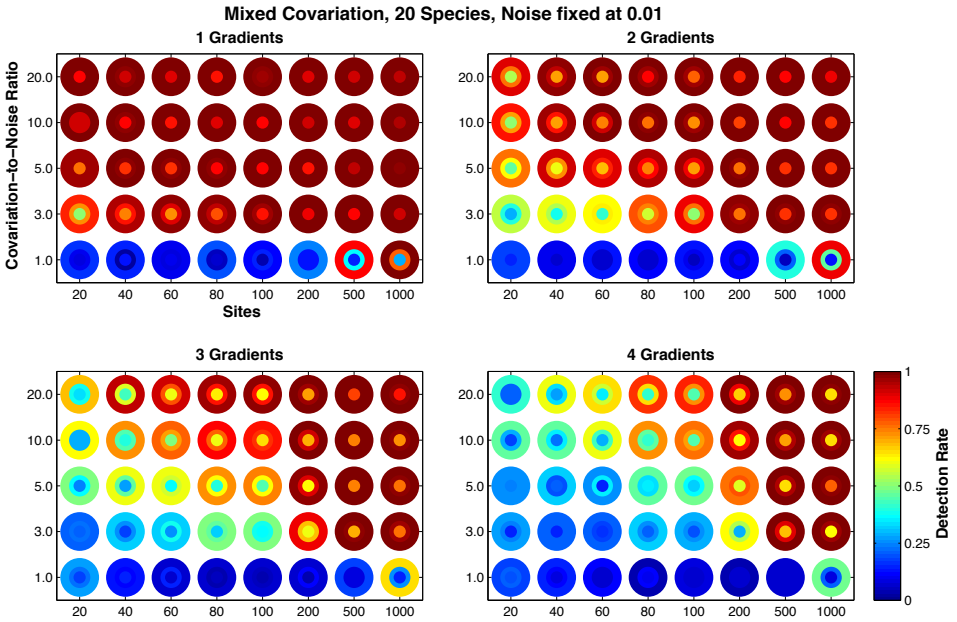

Figure B.1: Detection Rate: Mixed Community Covariation. Noise: 0.01, R^2_{avg} : 0.94.

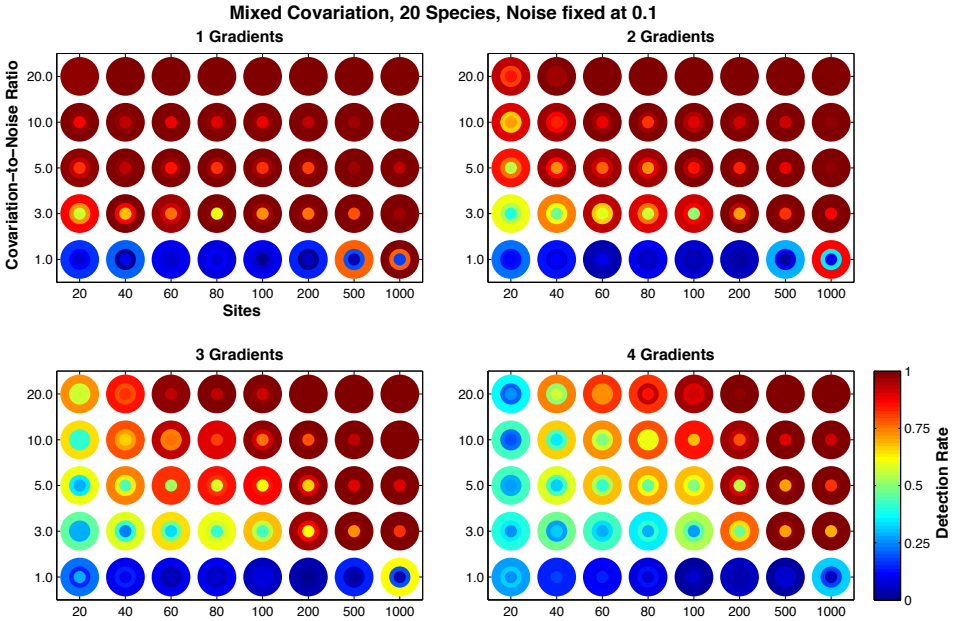

Figure B.2: Detection Rate: Mixed Community Covariation. Noise: 0.1, R^2_{avg} : 0.60.

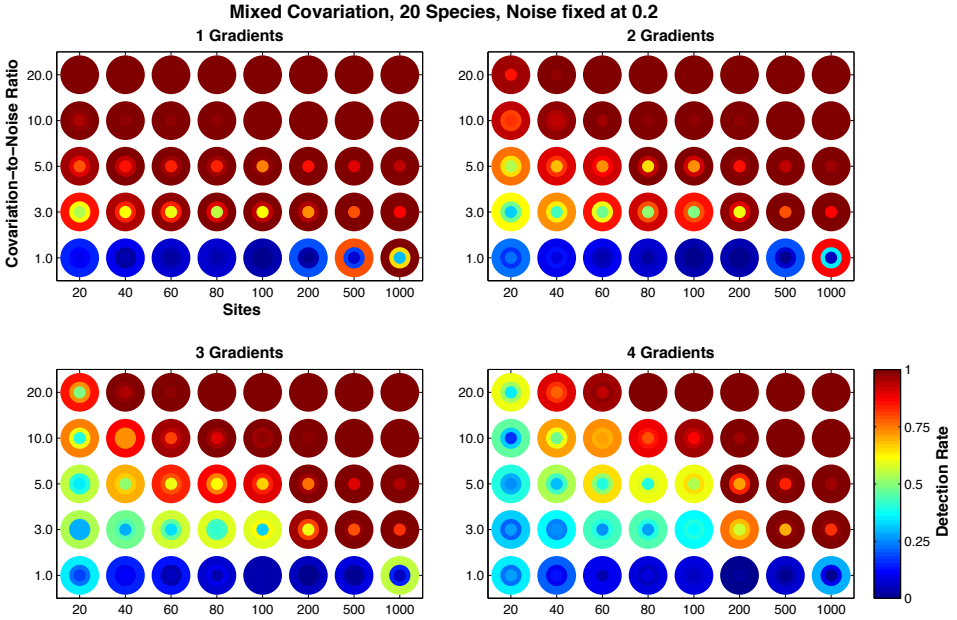

Figure B.3: Detection Rate: Mixed Community Covariation. Noise: 0.2, R^2_{avg} : 0.40.

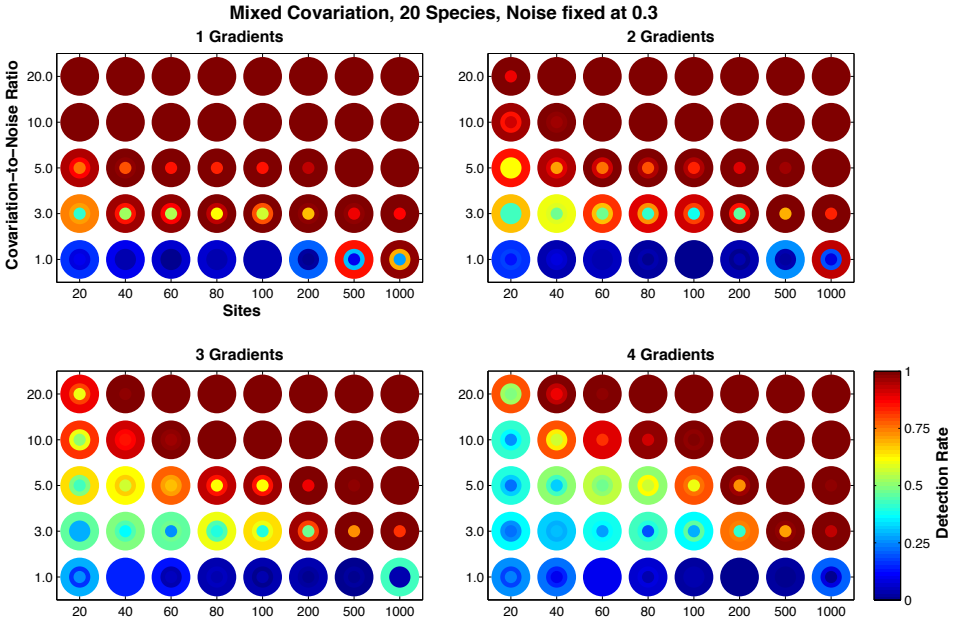

Figure B.4: Detection Rate: Mixed Community Covariation. Noise: 0.3, R^2_{avg} : 0.27.

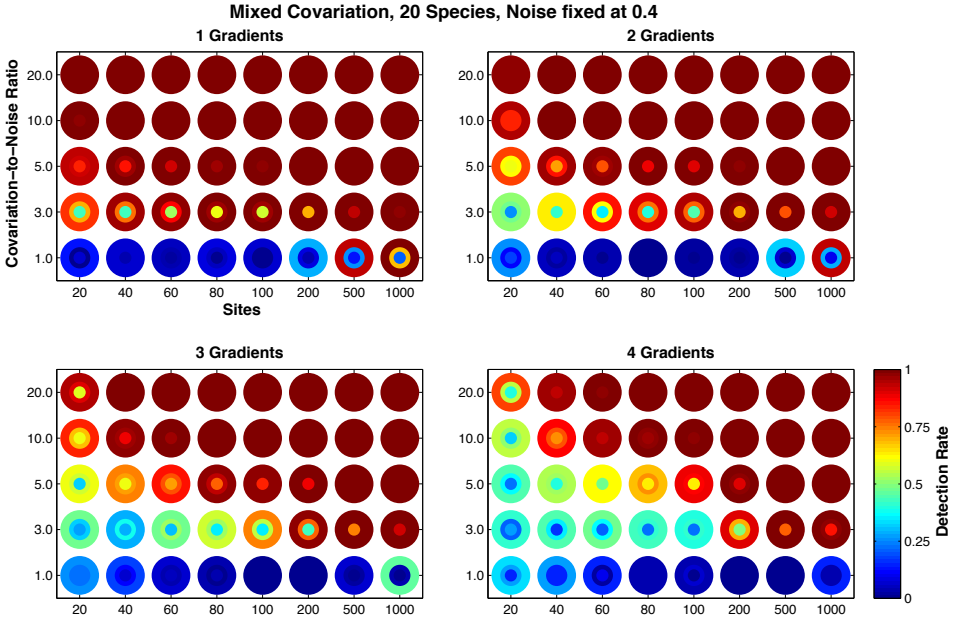

Figure B.5: Detection Rate: Mixed Community Covariation. Noise: 0.4, R^2_{avg} : 0.19.

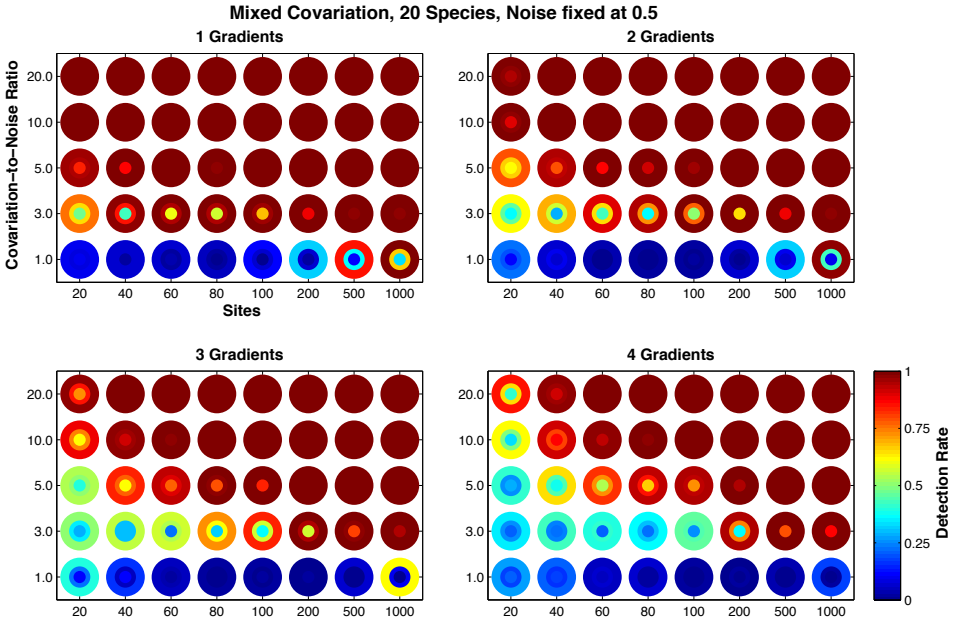

Figure B.6: Detection Rate: Mixed Community Covariation. Noise: 0.5, R^2_{avg} : 0.13.

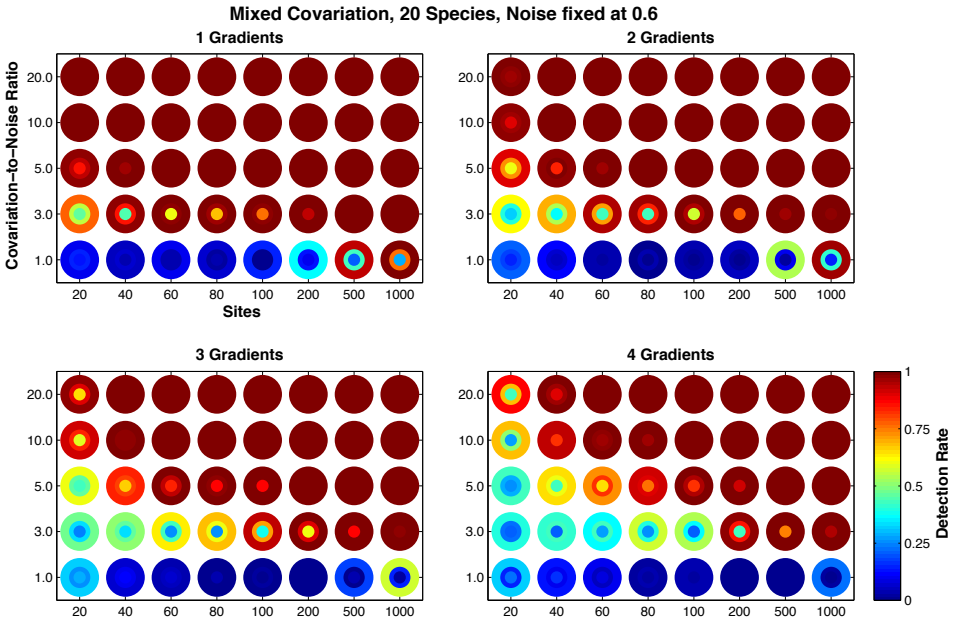

Figure B.7: Detection Rate: Mixed Community Covariation. Noise: 0.6, R^2_{avg} : 0.10.

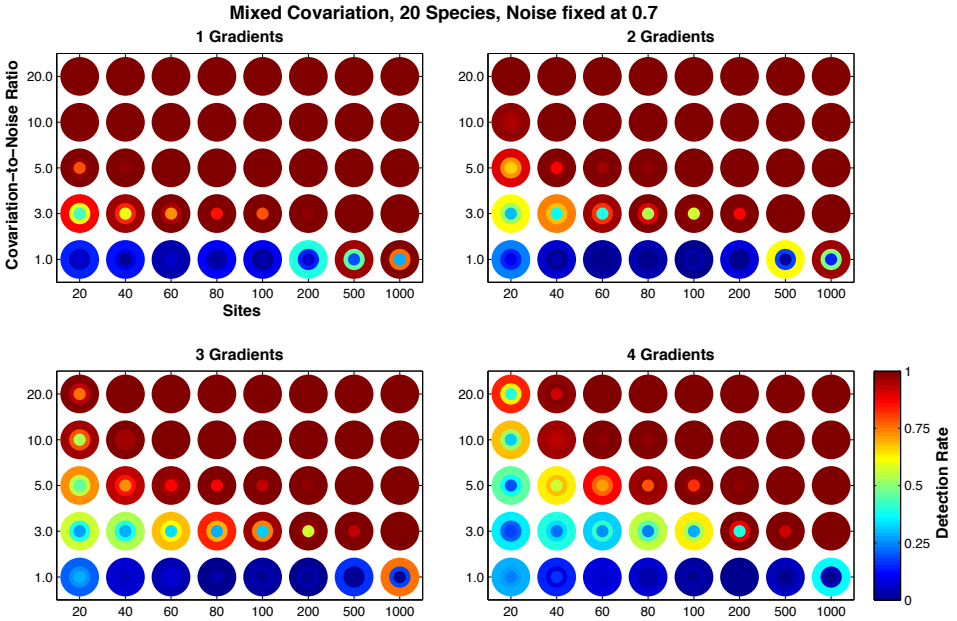

Figure B.8: Detection Rate: Mixed Community Covariation. Noise: 0.7, R^2_{avg} : 0.08.

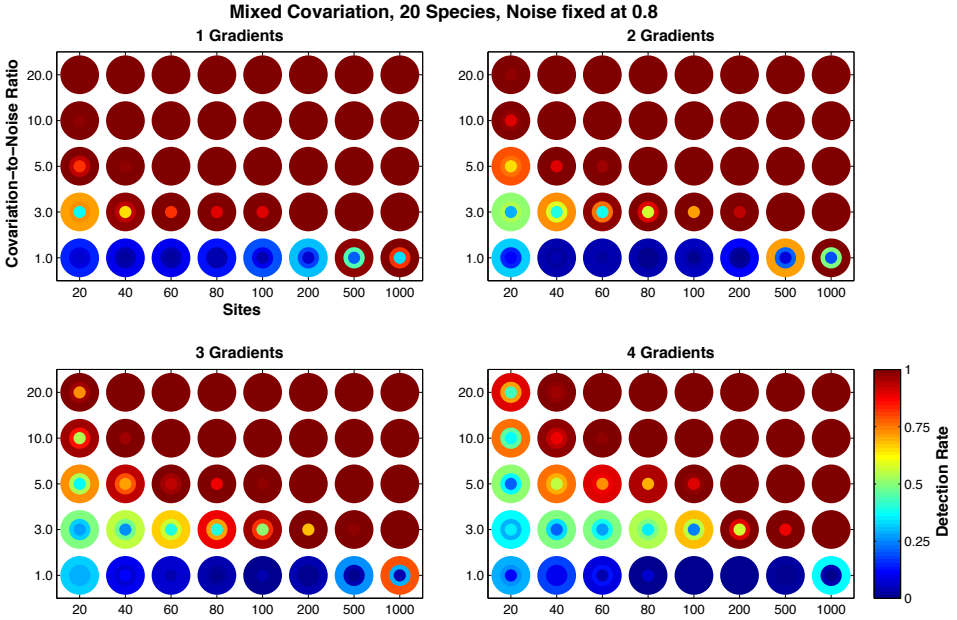

Figure B.9: Detection Rate: Mixed Community Covariation. Noise: 0.8, R^2_{avg} : 0.06.

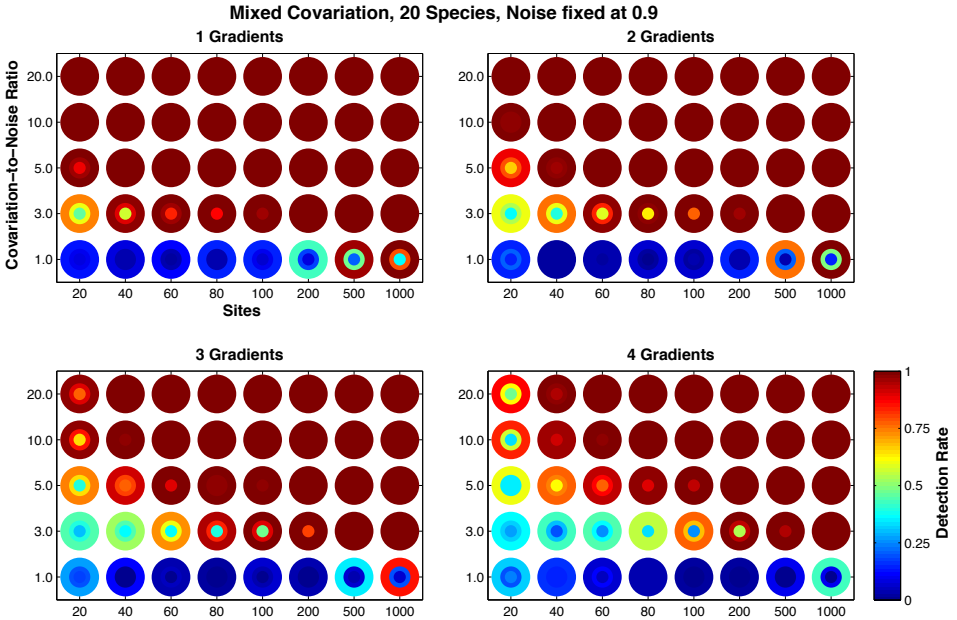

Figure B.10: Detection Rate: Mixed Community Covariation. Noise: 0.9, R^2_{avg} : 0.05.

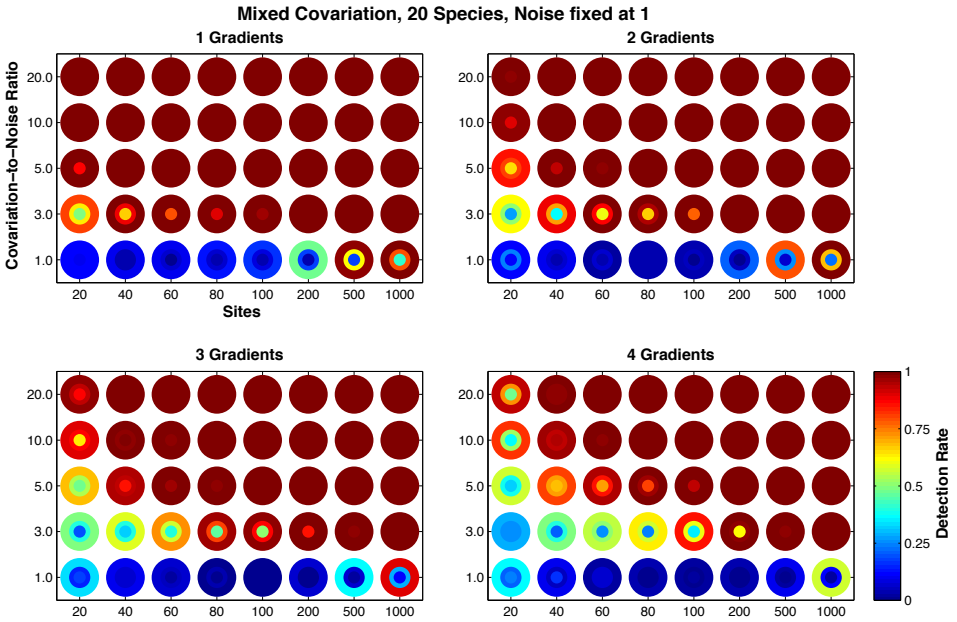

Figure B.11: Detection Rate: Mixed Community Covariation. Noise: 1, R^2_{avg} : 0.04.

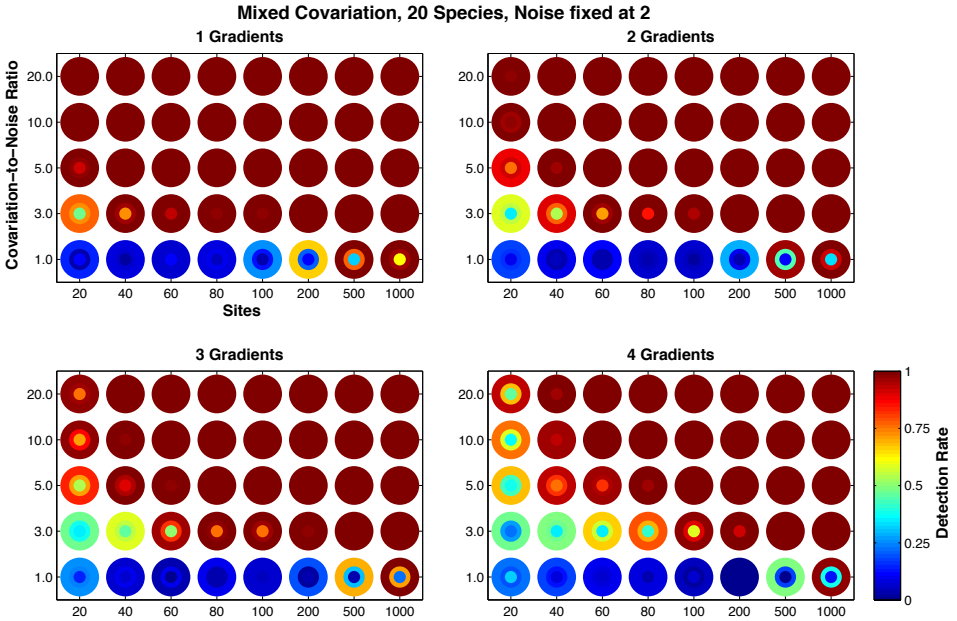

Figure B.12: Detection Rate: Mixed Community Covariation. Noise: 2, R^2_{avg} : 0.01.

Bibliography

[1] C. Braak and I. Prentice, *A Theory of Gradient Analysis*, ser. Advances in ecological research. Academic Press, 1988. [Online]. Available: http://books.google.com/books?id=OK9eQwAACAAJ

[2] J. Kuczynski, Z. Liu, C. Lozupone, D. McDonald, N. Fierer, and R. Knight, "Microbial community resemblance methods differ in their ability to detect biologically relevant patterns." *Nature methods*, Sep. 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20818378

[3] C. Rao, "THE USE AND INTERPRETATION OF PRINCIPAL COMPONENT ANALYSIS IN APPLIED RESEARCH," *Sankhyā: The Indian Journal of Statistics, Series A*, vol. 26, no. 4, pp. 329–358, 1964. [Online]. Available: http://www.jstor.org/stable/10.2307/25049339

[4] A. L. Van Den Wollenberg, "Redundancy analysis an alternative for canonical correlation analysis," *Psychometrika*, vol. 42, no. 2, pp. 207–219, 1977.

[5] B. H. McArdle and M. J. Anderson, "Fitting multivariate models to community data: a comment on distance-based redundancy analysis," *Ecology*, vol. 82, no. 1, pp. 290–297, 2001.

[6] N. J. Gotelli, "Null Model Analysis of Species Co-Occurrence Patterns," *Ecology*, vol. 81, no. 9, p. 2606, Sep. 2000. [Online]. Available: http://www.jstor.org/stable/177478?origin=crossref

[7] E. Connor and D. Simberloff, "The Assembly of Species Communities: Chance or Competition?" *Ecology*, vol. 60, no. 6, pp. 1132–1140, 1979. [Online]. Available: http://www.jstor.org/stable/10.2307/1936961

[8] N. J. Gotelli and W. Ulrich, "The empirical Bayes approach as a tool to identify non-random species associations." *Oecologia*, vol. 162, no. 2, pp. 463–77, Feb. 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19826839

[9] O. Ovaskainen, J. Hottola, and J. Siitonen, "Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions." *Ecology*, vol. 91, no. 9, pp. 2514–21, Sep. 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20957941

[10] M. Begon, J. Harper, and C. Townsend, *Ecology*. Blackwell Science, 1998. [Online]. Available: http://books.google.com/books?id=GXFDMgEACAAJ

[11] P. Legendre and L. Legendre, *Numerical Ecology*, ser. Developments in Environmental Modelling. Elsevier Science, 2012. [Online]. Available: http://books.google.com/books?id=6ZBOA-iDviQC

[12] E. Pielou, *The Interpretation of Ecological Data: A Primer on Classification and Ordination*, ser. A Wiley-Interscience publication. Wiley, 1984. [Online]. Available: http://books.google.com/books?id=aUaJoxiyjJoC

[13] P. Legendre and M. Anderson, "Distance based redundancy analysis: Testing Multispecies Responses in Multifactoral Ecological Experiments," *Ecological Monographs*, vol. 69, no. 1, pp. 1–24, 1999. [Online]. Available: http://www.esajournals.org/doi/abs/10.1890/0012-9615(1999)069%5B0001:DBRATM%5D2.0.CO;2

[14] J. Gower, "Some distance properties of latent root and vector used in multivariate analysis," *Biometrika*, vol. 53, no. 3, pp. 325–338, 1966. [Online]. Available: http://biomet.oxfordjournals.org/content/53/3-4/325.short

[15] J. Kruskal, "Nonmetric Multidimensional Scaling: A numerical Method," *Psychometrika*, vol. 29, no. 2, pp. 115–129, 1964. [Online]. Available: http://www.springerlink.com/index/TJ18655313945114.pdf

[16] N. J. D. NAGELKERKE, "A note on a general definition of the coefficient of determination," *Biometrika*, vol. 78, no. 3, pp. 691–692, 1991. [Online]. Available: http://biomet.oxfordjournals.org/content/78/3/691.abstract

[17] P. R. Peres-Neto, P. Legendre, S. Dray, and D. Borcard, "Variation partitioning of species data matrices: estimation and comparison of fractions." *Ecology*, vol. 87, no. 10, pp. 2614–25, Oct. 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17089669

[18] A. V. D. Wollenberg, "Redundancy Analysis An alternative for canonical correlation analysis," *Psychometrika*, vol. 42, no. 2, pp. 207–219, 1977. [Online]. Available: http://link.springer.com/article/10.1007/BF02294050

[19] C. T. Braak, "Canonical Correspondence Analysis : A New Eigenvector Technique for Multivariate Direct Gradient Analysis," *Ecology*, vol. 67, no. 5, pp. 1167–1179, 1986. [Online]. Available: http://www.esajournals.org/doi/abs/10.2307/1938672

[20] J. Oksanen and P. R. Minchin, "Continuum theory revisited: what shape are species responses along ecological gradients?" *Ecological Modelling*, vol. 157, no. 2, pp. 119–129, 2002.

[21] M. Hill and H. Gauch, "Detrended Correspondence Analysis: An imporved ordination technique," *Plant Ecology*, vol. 42, no. 1, pp. 47–58, 1980. [Online]. Available: http://www.springerlink.com/index/R1L713162404J437.pdf

[22] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. a. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. a. Vogt, G. a. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg, "Genome sequencing in microfabricated high-density picolitre reactors." *Nature*, vol. 437, no. 7057, pp. 376–80, Sep. 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16056220

[23] T. G. Martin, B. A. Wintle, J. R. Rhodes, P. M. Kuhnert, S. A. Field, S. J. Low-Choy, A. J. Tyre, and H. P. Possingham, "Zero tolerance ecology: improving ecological inference by modelling the source of zero observations," *Ecology Letters*, vol. 8, no. 11, pp. 1235–1246, 2005. [Online]. Available: http://dx.doi.org/10.1111/j.1461-0248.2005.00826.x

[24] C. Rao, *Linear Statistical Inference and Its Applications*, ser. Wiley Series in Probability and Statistics. Wiley, 1973. [Online]. Available: http://books.google.com/books?id=g02fNa9Mfn8C

[25] T. Amemiya, "TOBIT MODELS: A SURVEY," *Journal of econometrics*, vol. 24, no. 81, pp. 3–61, 1984. [Online]. Available: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:TOBIT+MODELS:+A+SURVEY#0

[26] M. Wolynetz, "Algorithm as 139: Maximum likelihood estimation in a linear model from confined and censored normal data," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 2, pp. 195–206, 1979.

[27] P. W. Holland and R. E. Welsch, "Robust regression using iteratively reweighted least-squares," *Communications in Statistics - Theory and Methods*, vol. 6, no. 9, pp. 813–827, 1977. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/03610927708827533

[28] F. Pukelsheim, "The three sigma rule," *The American Statistician*, vol. 48, no. 2, pp. 88–91, 1994. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/00031305.1994.10476030

[29] M. Leibold and G. Mikkelson, "Coherence, species turnover, and boundary clumping: elements of meta-community structure," *Oikos*, vol. 2, no. October 2001, pp. 237–250, 2002. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1034/j.1600-0706.2002.970210.x/full

[30] M. C. Horner-Devine, J. M. Silver, M. a. Leibold, B. J. M. Bohannan, R. K. Colwell, J. a. Fuhrman, J. L. Green, C. R. Kuske, J. B. H. Martiny, G. Muyzer, L. Ovreå s, A.-L. Reysenbach, and V. H. Smith, "A comparison of taxon co-occurrence patterns for macro- and microorganisms." *Ecology*, vol. 88, no. 6, pp. 1345–53, Jun. 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17601127

[31] S. J. Presley, C. L. Higgins, and M. R. Willig, "A comprehensive framework for the evaluation of metacommunity structure," *Oikos*, vol. 119, no. 6, pp. 908–917, Apr. 2010. [Online]. Available: http://doi.wiley.com/10.1111/j.1600-0706.2010.18544.x

[32] N. J. Gotelli and W. Ulrich, "Statistical challenges in null model analysis," *Oikos*, vol. 121, no. 2, pp. 171–180, Feb. 2012. [Online]. Available: http://doi.wiley.com/10.1111/j.1600-0706.2011.20301.x

[33] N. J. Gotelli and G. R. Graves, *Null models in ecology*. Smithsonian Institution Press Washington, DC, 1996.

[34] B. Manly, *Randomization, Bootstrap and Monte Carlo Methods in Biology, Second Edition*, ser. Chapman & Hall texts in statistical science series. Taylor & Francis, 1997. [Online]. Available: http://books.google.com/books?id=Q5nPYgfX9s8C

[35] E. Edgington and P. Onghena, *Randomization Tests, Fourth Edition*, ser. Statistics: A Series of Textbooks and Monographs. Taylor & Francis, 2007. [Online]. Available: http://books.google.com/books?id=qior0VqrhugC

[36] L. Stone and A. Roberts, "The checkerboard score and species distributions," *Oecologia*, vol. 1990, no. 85, pp. 74–79, 1990. [Online]. Available: http://link.springer.com/article/10.1007/BF00317345

[37] M. Dwass, "Modified randomization tests for nonparametric hypotheses," *The Annals of Mathematical Statistics*, vol. 28, no. 1, pp. 181–187, 1957.

[38] F. Marriott, "Barnard's monte carlo tests: How many simulations?" *Applied Statistics*, pp. 75–77, 1979.

[39] M. D. Moran, "Arguments for rejecting the sequential bonferroni in ecological studies," *Oikos*, vol. 100, no. 2, pp. 403–405, 2003.

[40] J. Neyman and E. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, no. 1933, pp. 289–337, 1933. [Online]. Available: http://link.springer.com/chapter/10.1007/978-1-4612-0919-5_6

[41] J. Shaffer, "Multiple hypothesis testing," *Annual review of psychology*, vol. 46, pp. 561–84, 1995. [Online]. Available: http://www.annualreviews.org/doi/pdf/10.1146/annurev.ps.46.020195.003021

[42] J. Aitchison, "A concise guide to compositional data analysis," *CDA Workshop, Girona*, 2003. [Online]. Available: http://www.leg.ufpr.br/lib/exe/fetch.php/pessoais:abtmartins:a_concise_guide_to_compositional_data_analysis.pdf

[43] P. Legendre and E. Gallagher, "Ecologically meaningful transformations for ordination of species data," *Oecologia*, vol. 129, no. 2, pp. 271–280, Oct. 2001. [Online]. Available: http://link.springer.com/10.1007/s004420100716

[44] C. Rao, "A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance," *Questiió: Quaderns d'Estadística, Sistemes, Informatica ...*, 1995. [Online]. Available: http://dialnet.unirioja.es/servlet/articulo?codigo=2362689

[45] F. G. Blanchet, P. Legendre, and D. Borcard, "Forward selection of explanatory variables." *Ecology*, vol. 89, no. 9, pp. 2623–32, Sep. 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18831183

[46] F. E. Clements, *Plant succession: an analysis of the development of vegetation.* Carnegie Institution of Washington, 1916, no. 242.

[47] H. Gleason, "The Individualistic Concept of the Plant Association," *Bulletin of the Torrey Botanical Club*, vol. 53, no. 1, pp. 7–26, 1926. [Online]. Available: http://www.jstor.org/stable/10.2307/2479933

[48] L. Wu, L. M. Feinstein, O. Valverde-Barrantes, M. W. Kershner, L. G. Leff, and C. B. Blackwood, "Placing the effects of leaf litter diversity on saprotrophic microorganisms in the context of leaf type and habitat." *Microbial ecology*, vol. 61, no. 2, pp. 399–409, Feb. 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20972562

[49] W. Liu and T. Marsh, "Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA," *Applied and environmental microbiology*, vol. 63, no. 11, p. 4516, 1997. [Online]. Available: http://aem.asm.org/content/63/11/4516.short

[50] J. M. Janda and S. L. Abbott, "16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls." *Journal of clinical microbiology*, vol. 45, no. 9, pp. 2761–4, Sep. 2007. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2045242&tool=pmcentrez&rendertype=abstract

[51] H. a. Erlich, "Polymerase chain reaction," *Journal of Clinical Immunology*, vol. 9, no. 6, pp. 437–447, Nov. 1989. [Online]. Available: http://link.springer.com/10.1007/BF00918012