Principal Response Curve Analysis of Arthropod Community Abundance Data with Sparse Subsets

Short Title: Principal Response Curve Analysis of Sparse Abundance Data

Changjian Jiang, C. R. Brown, P. Asiimwe, Chen Meng, Adam W. Schapaugh
Bayer Crop Science
700 Chesterfield Parkway West
Chesterfield, Missouri, 63017-1732, USA

Abstract

Principal response curve (PRC) analysis was applied to an assessment of the ecological impact of the genetically-modified (GM), insect-resistant, cotton MON 88702 on predatory *Hemiptera* communities in the field. The field community was represented by ten taxa collected ten times across the season at six sites, in which individual taxa were not observed in at least 25% of the time (unique site x collection combinations). These complete absences and those nearly so, called sparse subsets of the data in this investigation, were the result of geoclimatic and seasonal variations, which are both independent of the treatment effect for which the PRC analysis is intended. If the sparse subsets were included in the analysis, the treatment effect would be underestimated. Here, a modified analysis is proposed to remove those sparse subsets and to be performed on the incomplete data. In the application to MON 88702, four components (PRC1-4) were significant at the 5% level by the modified method, when more than 50% of the data were excluded due to no- or low responses, and five (PRC1-5) by the classical method. While PRC1-2 was highly consistent between two methods, PRC3-5 was largely different because of sparse subsets of the data. Differences in results between two methods demonstrate that excluding sparse subsets prevented the bias in the estimation of the treatment effect and the relationship with the community from confounding with the environmental variation that caused the sparse data. In this regard, the modification should be considered as a supplement of the classical PRC analysis and recommended when abundance data have sparse subsets.

Key words: Principal response curve analysis; Reduced rank regression; GM crop safety assessment; Analysis of arthropod community abundance; Sparse abundance data.

Introduction

Principal response curve (PRC) analysis is a multivariate statistical method for the ecological assessment of stress effects on a community across temporary/spatial intervals. A large proportion of the application has been on abundance data by transformation (see Paliy and Shankar 2016 for a review). The analysis can reveal inter-related responses of the treatment among species representative of the community, so that, the derived relationship can be used to order and compare the community responses in a reduced dimensional space (Van den Brink and Ter Braak 1998; Ter Braak and Smilauer 1998b, 2015). The algorithm of the analysis is a least-square based multivariate analysis of variance (MANOVA). The implementation of the analysis has been relying on the software CANOCO (Ter Braak and Smilauer 1998a; Smilauer and Leps 2014).

Often, individual species are absent or rare at given intervals, which will be called sparse subsets of the data in this investigation, due to large environmental variations independent of the treatment effect for which the PRC analysis is intended. Firstly, these sparse subsets, if analyzed as part of complete data, would underestimate the main treatment effect as an average over intervals due to lack of information for the treatment comparison. Secondly, sparse subsets would inflate interactions between the treatment and intervals with and without sparse data thus confound the treatment effect with the environmental variation that caused sparse data. The same would be true for covariances between species. Despite the generalized linear model likely being a better solution for sparse subsets, the classical PRC analysis by transformation was considered as computationally attractive in the ordination context due to a large number of species, and therefore has been the main option for the analysis of abundance data (Naranjo 2005; Ter Braak and Smilauer 2015; Auber et al. 2017). Nonetheless, the objection of a transformation has been well received in the literature (O'Hara and Kotze 2010). To a large extent, these objections were results of the biased estimation due to completely sparse data or a fraction of sparse subsets which hold true for both univariate and multivariate analyses.

In this investigation, a modification was proposed for the PRC analysis of abundance data to remove sparse subsets using a predetermined threshold (criterion). Although the remaining data are unsuitable for MANOVA by the classical PRC analysis, the modified method was proposed to estimate the multivariate treatment variations by the univariate analysis of variance (ANOVA) separately for each species and analysis of covariance (ANCOVA) for each pair of species. The modified method will become the classical analysis if no sparse subset presents. The calculation of the modified method can be performed by any statistical software e.g., SAS (SAS 2012). An application to a motivating example of a genetically-modified, insect-resistant, cotton showed consistent estimates of the first two canonical components between two methods, but substantial differences in other components because of sparse subsets. The modified method prevented the bias in the estimation of the treatment effect on the arthropod community from confounding with the environmental variation that caused the sparse data. Hence, the modified method should be recommended for the PRC analysis of abundance data when sparse subsets present.

In the following sections, at first the motivating example is described, and followed by introductions of the classical PRC analyses. Then, the modification was proposed and illustrated by the example. Additional discussions can be found in the discussion section.

A motivating example: MON 88702 ecological safety data

MON 88702 is a genetically-modified (GM) cotton developed by Bayer Crop Science with an insect-resistant trait targeted a *hemipteran* pest Lygus (Bachman et al. 2017). A field trial was conducted in 2018 at six sites representative of U.S. cotton growing regions each with a randomized complete block design of three blocks (see Asiimwe et al 2021 for details). The objective of the trial is an ecological assessment of the impact of MON 88702 on the abundance of the predatory *Hemiptera* community under the field condition with the traditional management including insecticide application. Five treatments consist of combinations of (multiple) insecticide applications and cotton varieties (GM variety MON 88702 and a near-isogenic conventional control DP393) abbreviated as: C2 = DP393 with a conventional broad-spectrum insecticide, C1 = DP393 with minimal or selective insecticides, T1 = MON 88702 with minimal or selective insecticides, C0 = DP393 with no insecticide, T0 = MON 88702 with no insecticide. Abundances of ten taxa representative of the community were collected ten times over the course of the season and analyzed in this investigation.

A summary of the data was listed in Table 1. The table shows large variations of abundance among sites/collections due to geoclimatic and seasonal differences. Out of 600 (10 taxa x 6 sites x 10 collections) combinations (each with 15 responses of five treatments and three replicates), 150 (25%) have zero abundance. When analyzed as complete data, these zero-abundance subsets will underestimate the main treatment effect due to averaging. The same would be true for subsets with low abundance as will be shown next.

Table 1. MON 88702 ecological abundance data: Mean counts and numbers of collections with non-zero abundance by arthropod taxon and site

Taxa	**Mean Count (# of collections with capture > 0) by Site**						**Mean**
	AZMA	**AZYU**	**LACH**	**MSGV**	**NCRC**	**TXUV**	
Aphids	0.2 (7)	0.0 (3)	203.9 (10)	85.7 (10)	34.0 (10)	1.5 (7)	54.2 (7.8)
Cotton Flea hoppers	4.4 (10)	4.3 (10)	0.0 (1)	0.1 (1)	0.2 (10)	13.9 (10)	3.8 (7.0)
Geocoris	6.0 (10)	8.7 (10)	1.1 (10)	10.4 (10)	2.3 (10)	3.4 (10)	5.3 (10)
Lygus	3.4 (10)	2.0 (10)	2.4 (9)	5.4 (10)	3.9 (10)	0.0 (3)	2.9 (8.7)
Nabis	0.3 (4)	1.1 (10)	0.0 (3)	0.5 (10)	0.5 (10)	0.0 (2)	0.4 (6.5)
Orius	5.4 (10)	5.7 (10)	2.5 (10)	4.2 (10)	7.1 (10)	91.3 (10)	19.4 (10)
Predatory Stink bugs	0.0 (0)	0.1 (2)	0.2 (6)	1.2 (8)	0.3 (7)	0.0 (0)	0.3 (3.8)
Stink bugs	0.9 (8)	0.3 (9)	0.2 (5)	1.3 (9)	1.0 (9)	0.3 (6)	0.7 (7.7)
Whiteflies	74.0 (9)	57.1 (10)	0.2 (6)	34.9 (7)	0.6 (8)	158.7 (9)	54.3 (8.2)
Zelus	2.7 (10)	0.7 (8)	0.2 (3)	0.0 (1)	0.0 (0)	1.7 (10)	0.9 (5.3)
Average	9.7 (7.8)	8.0 (8.2)	21.1 (6.3)	15.1 (7.6)	5.0 (8.4)	25.8 (6.7)	14.1 (7.5)

As abundance varying as wide as those in Table 1, to see how assumptions of a linear model analysis would be satisfied by a transformation, a simulation study was performed, and results were presented in Fig 1. Let *x* be the observed count, and $y = \sqrt{x}$ and $y = log\ (x + 1)$ representing two popular transformations in practice (St-Pierre et al. 2018).

Figure 1 is the graphical representations of sample mean difference (Diff) and standard error (SE) and t ratio (= Diff/SE) as functions of the expected mean count with and without transformation. Two random samples of $n = 10$ were generated from two negative binomial distributions with mean differing by 25% and average of two means ranging $(0.1{\sim}10)$ and variance of each distribution $\sigma^2 = \mu + a\mu^2$ as $a = (0, 0.1, 0.2)$. Scales include count (Count), square root (SQRT), and logarithm (Log). Each point represents a mean of 10000 replications.

Fig 1 provides convincing evidence of the roles of a transformation in a linear model analysis of abundance data which were generated from a generalized distribution for count (despite slight differences in performance by two transformations). Without transformation, both mean (the top plots) and standard error (the middle plots) of the sample difference are increasing functions of the average of two distribution means, and after transformation, both mean difference and the standard error are effectively stabilized over a wide range of abundance except for the average approximately < 1. Therefore, assumptions of linearity (or additivity of the treatment effect) and homogeneity (of variance across intervals) of a linear model are mostly achieved by both transformations if the average count is approximately > 1. These results indeed support Warton's remark (2005) "Surprisingly, transformed least squares appeared to fit data about as well as" a generalized linear model. However, Fig 1 also shows the poor performance of transformation when the mean count is low (O'Hara and Kotze 2010). Most samples from those distributions with low means could be called sparse data as defined in this investigation, which provide little information for the treatment comparison as shown by low values of the t ratio (the bottom plots). Therefore, a threshold of mean count < 1.0 was applied for defining the sparse subsets of abundance data in this investigation. In spite a somewhat arbitrary definition, different thresholds in a range of (0.5 ~ 1.0) were not shown to make substantial differences likely due to the large environmental variation as the main source of sparse data. The application of the criterion will be discussed further in the analysis of MON 88702 data.

For the analysis of MON 88702 data, despite those sparse subsets as shown in Table 1, the PRC analysis by transformation is likely the only option due to multiple taxa. Thus, a modified analysis was proposed next. At first, a brief introduction was provided for the classical method. Then, the modification was described according to steps of a classical analysis.

PRC analysis of abundance data with sparse subsets

RDA of Davies and Tso: The following multivariate model was assumed by Davies and Tso (1982) for the RDA.

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{M} + \boldsymbol{E} \qquad (1)$$

where $\boldsymbol{Y}(n \times p)$ consists of p responses observed from n samples, $\boldsymbol{X}(n \times h)$ represents the design matrix, $\boldsymbol{M}(h \times p)$ for the treatment effect, and $\boldsymbol{E}(n \times p)$ for the residual. RDA assumes that there exists a factorization $\boldsymbol{M_s} = \boldsymbol{A}(h \times s)\boldsymbol{B}(s \times p)$ with $s < min(p, h)$ which could provide a reduced model equivalent to (1) with the rank of $\boldsymbol{M_s}$ spend by s canonical components less than the rank of $\boldsymbol{M}$. That is, a reduced rank model assumes that

$$\boldsymbol{Y} = \boldsymbol{FB} + \Delta + \boldsymbol{E} \tag{2}$$

where $\boldsymbol{F} = \boldsymbol{XA}$ represents a set of new (or latent) independent variables, and $\boldsymbol{B}$ is a new set of coefficients, and Δ (a part of $\boldsymbol{XM}$ in (1)) is assumed to be comparable with $\boldsymbol{E}$.

Let $\boldsymbol{SS}_Y$, $\boldsymbol{SS}_{\hat{Y}}$, and $\boldsymbol{SS}_{\hat{Y}_s}$ denote sums of squares and cross products (SSCP) of the observed and the estimated responses from fitting the models (1) and (2). An estimate $\widehat{\boldsymbol{Y}}_{\boldsymbol{s}} = \widehat{\boldsymbol{F}}\widehat{\boldsymbol{B}} = \boldsymbol{X}\widehat{\boldsymbol{M}}_{\boldsymbol{s}}$ under (2) can be obtained from a multivariate analysis of variance (MONOVA) and the single value decomposition (SVD) of $\boldsymbol{SS}_{\hat{Y}}$ (detailed in the following sections) through

$$\boldsymbol{SS}_Y = \boldsymbol{SS}_{\hat{Y}} + \boldsymbol{SS}_e = \boldsymbol{SS}_{\hat{Y}_s} + \boldsymbol{SS}_{\hat{\Delta}} + \boldsymbol{SS}_e \tag{3}$$

where $\boldsymbol{SS}_e$ is the residual under (1), and the hypothesis H_0: $\boldsymbol{M} = \boldsymbol{M}_{\boldsymbol{s}} = \boldsymbol{AB}$ can be statistically tested by comparing $\boldsymbol{SS}_{\hat{\Delta}}$ with $\boldsymbol{SS}_e$.

The classical PRC analysis: The PRC analysis was developed by Van den Brink and Ter Braak (1998) and Ter Braak and Smilauer (1998b) for longitudinal ecological data with the model

$$\boldsymbol{Y} = \boldsymbol{XM} + \boldsymbol{ZD} + \boldsymbol{E} \tag{4}$$

where $\boldsymbol{M}$ is for the treatment effect as in (1), and $\boldsymbol{D}$ is for those of environmental factors including (temporary/spatial) intervals. The PRC analysis is a partial RDA interested only in the treatment effect $\boldsymbol{Y}|\boldsymbol{Z} = \boldsymbol{XM}$, including main effects and interactions of the treatment with the interval as the name PRC suggests. More importantly, a Monte Carlo permutation test was proposed for the significance of each canonical component of $\boldsymbol{M}_{\boldsymbol{s}}$ and implemented in CANOCO (Ter Braak and Smilauer 1998a, and Smilauer and Leps 2014). The test can separate the true components from random residuals of the same order (Legendre and Ter Braak 2011).

Both above procedures rely on the least-squares based MANOVA of complete data of p responses from n samples.

Modified PRC analysis of abundance data with sparse subsets: In this section, MON 88702 data were used as an example for abundance data with sparse subsets. Let y_{kijlm} be a transformed response of the k^{th} taxon from the i^{th} site, the j^{th} block, the m^{th} collection with the l^{th} treatment. A univariate form of (4) for the k^{th} taxon (except the covariance structure) can be expressed as

$$y_{kijlm} = \mu_k + s_{ki} + r_{kij} + a_{kl} + t_{kim} + (as)_{kil} + (at)_{kilm} + e_{kijlm} \qquad (5)$$

where μ_k is the taxon mean, a_{kl}, s_{ki}, r_{kij}, t_{kim} are main effects of the treatment, site, replicate and collection within site, $(as)_{kil}$ and $(at)_{kilm}$ are the corresponding interactions, and e_{kijlm} is the residual. If y_{kijlm} is in sparse subsets, the modified method was proposed to exclude y_{kijlm} from the analysis which was described next in steps of a classical PRC analysis.

At first, with the exclusion of sparse subsets, the estimation of $\boldsymbol{SS}_{\hat{Y}|Z}$ of the classical PRC analysis must be modified due to incomplete responses. Despite the exclusion, however, the treatment is still balanced with the replicate in the remaining collections. Let $\bar{y}_{kilm}$ and $\bar{y}_{ki.m}$ be means of the l^{th} treatment and over all treatments, respectively, at a given site and collection. All means are over replicates with the subscript omitted. Elements of $\boldsymbol{SS}_{\hat{Y}|Z}$ can be calculated as

$$\left(\boldsymbol{SS}_{\hat{Y}|Z}\right)_{k_1k_2} = n_r \sum_i \sum_l \sum_m \left(\bar{y}_{k_1ilm} - \bar{y}_{k_1i.m}\right)\left(\bar{y}_{k_2ilm} - \bar{y}_{k_2i.m}\right) \qquad (6)$$

Let $\boldsymbol{SS}_a$, $\boldsymbol{SS}_{as}$, and $\boldsymbol{SS}_{at(s)}$ be SSCPs of a_{kl}, $(as)_{kil}$ and $(at)_{kilm}$ in (5), which can be calculated in a similar way. Then, $\boldsymbol{SS}_{\hat{Y}|Z}$ can be decomposed into or calculated alternatively as a sum as

$$\boldsymbol{SS}_{\hat{Y}|Z} = \boldsymbol{SS}_a + \boldsymbol{SS}_{as} + \boldsymbol{SS}_{at(s)} \qquad (7)$$

Let $c_{k_1k_2i}$ be the observed number of collections at the i^{th} site with both ${k_1}^{th}$ and ${k_2}^{th}$ taxa. Let n_a, n_s, n_r and n_c be numbers of the treatment, site, block, and collection per site for complete data. Let $\bar{y}_{ki..}$, $\bar{y}_{k.l.}$ and $\bar{y}_{kil.}$ be means of a given site and treatment, and $\bar{y}_{k\ldots}$ be the overall mean. The decomposition of (7) can be calculated as

$$\begin{cases} (\boldsymbol{SS}_a)_{k_1k_2} = n_r c_{k_1k_2} \sum_l \left(\bar{y}_{k_1.l.} - \bar{y}_{k_1\ldots}\right)\left(\bar{y}_{k_2.l.} - \bar{y}_{k_2\ldots}\right) \\ (\boldsymbol{SS}_{as})_{k_1k_2} = n_r \sum_i \sum_l c_{k_1k_2i} \left(\bar{y}_{k_1il.} - \bar{y}_{k_1i..} - \bar{y}_{k_1.l.} + \bar{y}_{k_1\ldots}\right) \\ \qquad\qquad\qquad\qquad \left(\bar{y}_{k_2il.} - \bar{y}_{k_2i..} - \bar{y}_{k_2.l.} + \bar{y}_{k_2\ldots}\right) \\ \left(\boldsymbol{SS}_{at(s)}\right)_{k_1k_2} = n_r \sum_i \sum_l \sum_m \left(\bar{y}_{k_1ilm} - \bar{y}_{k_1il.} - \bar{y}_{k_1i.m} + \bar{y}_{k_1i..}\right) \\ \qquad\qquad\qquad\qquad \left(\bar{y}_{k_2ilm} - \bar{y}_{k_2il.} - \bar{y}_{k_2i.m} + \bar{y}_{k_2i..}\right) \end{cases} \qquad (8)$$

where $c_{k_1k_2}$ is the total number of collections with both ${k_1}^{th}$ and ${k_2}^{th}$ taxa over all sites. Both (6) and (8) provide consistent estimate of $\boldsymbol{SS}_{\hat{Y}|Z}$ since each element is a least-square estimate though some off-diagonal elements could be zero. Notice that the modified method differs from the classical analysis in numbers of responses for $\bar{y}_{ki..}$, $\bar{y}_{k.l.}$, $\bar{y}_{ki.m}$, $\bar{y}_{kil.}$, and $\bar{y}_{k\ldots.}$.

In the second step, regardless of the exclusion of sparse subsets, $\boldsymbol{SS}_{\hat{Y}|Z}$ of (6) and (7) is non-negative definite and SVD of $\boldsymbol{SS}_{\hat{Y}|Z}$ can proceed as the classical analysis of Davies and Tso (1982). Let λ_i^2 and $\boldsymbol{u}_i$ be eigenvalues and eigenvectors of $\boldsymbol{SS}_{\hat{Y}|Z}$ with $\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_p^2$. A PRC analysis of rank s can be derived from

$$\boldsymbol{SS}_{\hat{\boldsymbol{Y}}_s|\boldsymbol{Z}} = \lambda_1^2 \boldsymbol{u}_1 \boldsymbol{u}_1' + \cdots + \lambda_s^2 \boldsymbol{u}_s \boldsymbol{u}_s' \tag{9}$$

where λ_i^2 represents the variation captured by the i^{th} canonical component (PRCi), $\boldsymbol{u}_i$ is the weight (or contribution) of each taxon, and $100 \cdot \lambda_i^2 / \sum_i \lambda_i^2$ is the percent variation captured.

For the third step, the significance test for each component of (9) remains the same as the classical analysis by a restricted permutation within each non-sparse combination of collections, blocks, and sites. The following pseudo-F statistic can be used for the test of the i^{th} component (also called a marginal test by Pierre et al. 2011). The P value of the test is the fraction of permutations with F values greater than the one from the analysis of the original data.

$$F = \frac{\lambda_i^2 / df_{hi}}{tr(\boldsymbol{SS}_e)/df_e}$$

where $\boldsymbol{SS}_e$ can be calculated as the difference between the total and individual SSCPs under (5) or as interactions of the replicate with the treatment and the collection within each site, df_{hi} and df_e are degrees of freedom of the i^{th} component and the residual and both are constants across permutations with no effects on the test. Comparing $\boldsymbol{M}(h \times p)$ of (1) with $\boldsymbol{M_s}(h \times p) = \boldsymbol{A}(h \times s)\boldsymbol{B}(s \times p)$ in (2), the number of parameters is reduced by $(p - s)(h - s)$ (Davies and Tso 1982). With sparse subsets excluded, h becomes $\tilde{h} = (n_a - 1)\sum_k (\sum_i c_{kki} - 1) / p$ on average over taxa. For $s = 2$, $df_{h1} = p + \tilde{h} - 1,\ df_{h2} = p + \tilde{h} - 3, df_e = p(n_r - 1)\tilde{h} + (n_r - 1)\sum_k \sum_i (c_{kki} - 1)$ with $c_{kki} \geq 1$. When PRC1 is significant, $(\hat{\boldsymbol{Y}}|\boldsymbol{Z})\boldsymbol{u}_1$ will be added to the covariates of $\boldsymbol{Z}$ and the permutation will be applied to the residual for the test of PRC2, and so on for additional components.

Though the PRC analysis is interested in the total treatment variation, each component of (9) can be decomposed under (5) into

$$\lambda_i^2 = \boldsymbol{u}_i'\boldsymbol{SS}_a\boldsymbol{u}_i + \boldsymbol{u}_i'\boldsymbol{SS}_{as}\boldsymbol{u}_i + \boldsymbol{u}_i'\boldsymbol{SS}_{at(s)}\boldsymbol{u}_i = \lambda_{ia}^2 + \lambda_{ias}^2 + \lambda_{iat(s)}^2 \tag{10}$$

For visualizing the canonical component of (10) by the ordination diagram, a univariate form of the reduced model similar to Van den Brink and Ter Braak (1999) can be expressed as

$$y_{kijlm}^* = b_{k1}\left(c_{ilm1}^{a} + c_{ilm1}^{as} + c_{ilm1}^{at(s)}\right) + \cdots + b_{ks}\left(c_{ilms}^{a} + c_{ilms}^{as} + c_{ilms}^{at(s)}\right) + e_{kijlm} \tag{11}$$

where y_{kijlm}^* represents the deviation of the l^{th} treatment from the mean of all treatments in a given site, replicate, and collection, b_k is the species weight, and c_{ilm}^{a}, c_{ilm}^{as}, and $c_{ilm}^{at(s)}$ are the main effects and interactions of each component, and e_{kijlm} is the combination of $(\Delta + \boldsymbol{E})$ in (2). Coefficients in (11) can be estimated from the difference in (8) and $\boldsymbol{u}_i$ in (10).

In the following analysis of MON 88702 data, the safety assessment of the GM trait relies on the comparison of the treatment effect across the community. While the modified method prevented the underestimation of the main treatment effect c_{ilm}^{a}, the classical PRC analysis, as results were compared between two methods, confound the treatment effect with the environmental variation which caused sparse subsets and inflated the estimation of the interactions c_{ilm}^{as} and $c_{ilm}^{at(s)}$.

PRC analyses of MON 88702 data

In this section, both the modified and the classical PRC analyses were applied to MON 88702 data. For the modified method, a threshold (criterion) of 1.0 was applied to each (taxon x collection) combination. That is, sparse subsets consist of not only these with no capture (25%) as shown in Table 1 but also those with 0 < mean < 1.0 (32.5%). Only 42.5% of the complete data exceeds the threshold with taxa Geocoris and Orius from all six sites, pest Lygus from five sites, and Nabis, Predatory stink bugs, and Stink bugs each from only one site. The observed counts were then transformed by the square root and scaled by the residual standard deviation (Ter Braak 1995) from the univariate analysis using (5). SAS procedures PROC MEANS were used for $\boldsymbol{SS}_{\hat{Y}|Z}$ in (6) and PROC IML was used for SVD in (9). PROC PLAN generated 1000 treatment permutations within each (site x collection x replicate) combination. The classical analysis was performed on the complete data with $\boldsymbol{SS}_{\hat{Y}|Z}$ estimated by PROC GLM option MONOVA. Results of two methods were compared in Table 2 and Fig 2.

Table 2. PRC analyses of MON 88702 data by the modified and the classical methods: Percentages of total treatment variation ($\%\lambda_i^2 = 100\,\lambda_i^2/\sum_i \lambda_i^2$) captured by each component and contribution of the main effect ($\%\lambda_{ia}^2 = 100\,\lambda_{ia}^2/\lambda_i^2$), and the statistical significance (P value).

PRC	**Modified PRC Analysis**		**Classical PRC Analysis**	
	$\%\lambda_i^2$ ($\%\lambda_{ia}^2$)	**P value**	$\%\lambda_i^2$ ($\%\lambda_{ia}^2$)	**P value**
PRC1	36.7 (44.9)	<0.001	29.5 (35.8)	<0.001
PRC2	17.3 (28.5)	<0.001	13.6 (22.6)	<0.001
PRC3	13.2 (15.9)	<0.001	12.1 (2.6)	<0.001
PRC4	10.4 (19.8)	<0.001	10.5 (1.0)	0.005
PRC5			8.1 (3.1)	0.035
Sum	77.6 (33.0)		73.8 (19.4)	

Table 2 lists four statistically significant components (PRC1-4) by the modified method and five (PRC1-5) by the classical method at the 5% level. The total variations account for 77.6% and 73.8% of the treatment variation over all taxa, and the modified method with fewer number of components captured a higher proportion of variation. The contribution of the main effect ($\%\lambda_{ia}^2$) are 33.0% and 19.4%, respectively, showing a substantial difference. PRC3-5 by the classical method consist almost entirely of the interaction. These results confirm that the classical analysis underestimates the main treatment effect and excluding sparse subsets by the modified method can prevent the confounding of the treatment effect with the environmental variation that caused the sparse data. Two methods were further compared next by the ordination diagram.

Fig 2 compares ordinations of the main treatment effect c_{ilm}^a (arrow) and the taxa weight b_k (dot) between two methods for PRC1-4. If two methods provided similar results, all arrows and dots would have lined up along the line ($x = y$). Indeed, PRC1-2 in Fig 2 are highly consistent between two methods. For PRC1, (C2, C1, T1) are all negative and (C0, T0) are positive demonstrating the depressing effect of the insecticide on most taxa indicated by the mostly positive weights; and in contrast for PRC2, (C2, C1, C0) are all positive and (T1, T0) are negative demonstrating the effect of the GM trait against the targeted pest Lygus indicated by a predominant weight 0.905 for the modified method and 0.791 for the classical method.

Fig 2. contains ordination diagrams of PRC analyses of MON 88702 data: Comparing the estimated main treatment effect (arrow)and taxa weight (dot) of PRC1-4 between the classical and modified methods. Labels were described in text for the treatment, and by initials of Table 1 for the taxa. To fit taxa weight b_k and the treatment effect c_{ilm}^{a} in the same plot, the treatment effect was multiplied by 0.75, and rescaled for PRC2-4 in reciprocally proportional to the square root of the variation relative to that of PRC1.

For PRC3-4 in Fig 2, however, large differences can be found between two methods. For example, three taxa ("N", "P", "S") each with data from only one site showed no effect as expected by the modified method but noticeable effects by the classical method. For PRC3 by the modified method, the treatment effect represents a contrast (C2, C1, T1) versus (C0, T0) for the effect of insecticide like PRC1 but differential effects among taxa indicated by b_k e.g., Lygus versus Cotton Flea hoppers. Similarly, PRC4 by the modified method represents a contrast (C1, T1) versus (C2, C0, T0), suggesting additional effect of the selective insecticide e.g., Orius versus Aphid which are mostly negligible in PRC1-3. However, PRC3-4 by the classical method show substantial weights for several taxa with mostly sparse data and provide no interpretation.

In summary, the classical PRC analyses of MON 88702 data demonstrated that a large proportion of sparse subsets in abundance data could substantially bias the estimation of the treatment effect. Excluding sparse subsets can provide estimates independent of the environmental variation which caused sparse data.

Discussion

Abundance data of an ecological community often include subsets of no- or low responses for species in certain spatial/temporary intervals regardless of the treatment, which have been called sparse subsets in this investigation. MON 88702 data is a typical example due to both geoclimatic and seasonal variations. It was shown in this investigation that sparse subsets, if included in PRC analysis by transformation, would underestimate the main treatment effect (due to averaging over intervals as shown in Fig 1) and inflate the interaction by confounding with the environment that caused sparse data. While a systematic treatment effect among species is the main interest of the PRC analysis under assumptions of the reduced rank model (2), the interactions with the environment which caused sparse data were largely random and should not be part of the PRC components. Excluding spare subsets can prevent this type of bias in estimating the treatment effects among species. Despite a possibly slight loss of information, a robust inference of the true relationship between the treatment effect and the community would justify the modification.

To define a criterion for sparse subsets in the MON 88702 example, alternative thresholds in the range (0.5 ~ 1.0) as well as the log-transformation were tried and only slight differences in results (not presented) were discovered. In practice, species abundance often varies in a wide range as is shown in Table 1 and a high proportion of sparse subsets with an average count in a narrow range e.g., (0.5 ~ 1.0) is not expected due to the large environmental variation. Hence, whether a threshold would be applied is much more important than which threshold should be chosen in preventing the estimation bias of the treatment effect except that a threshold too high may cause unnecessary loss of information.

The modified PRC analysis in this investigation estimates the multivariate treatment variation by the least-square based ANOVA separately for each species and ANCOVA for each pair of species. In addition to consistency of the estimation, the modified method is efficient, and easily implemented by most statistical software. Furthermore, if no sparse subset presents, the modified analysis will become the classical method. Hence, the modified method should be regarded as a supplement of the classical analysis of abundance and recommended whenever the proportion of sparse subsets are substantial, say > 10% of the complete data.

Acknowledgement

We are grateful to Dr. Zhenglei Guo for useful suggestions about the manuscript.

References

Asiimwe P., Brown C. R., Ellsworth P. C., Reisig D. D., Bertho L., Jiang C., Schapaugh A., Head G., Burzio L. (2021) Transgenic cotton expressing mCry51Aa2 does not adversely impact beneficial non-target Hemiptera in the field. Biological Control (in progress).

Auber A., Travers-Trolet M., Villanueva M. C., Ernande B. (2017) A new application of principal response curves for summarizing abrupt and cyclic shifts of communities over space. Ecosphere. 8(12): 1-22.

Bachman P. M., Ahmad A., Ahrens J. E., Akbar W., Baum J. A., Brown S., Clark T. L., Fridley J. M., Gowda A., Greenplate J. T., Jensen P. D., Mueller G. M., Odegaard M. L., Tan J., Uffman J. P., Levine S. L. (2017) Characterization of the Activity Spectrum of MON 88702 and the Plant-Incorporated Protectant Cry51Aa2.834_16. PLoS ONE 12(1): e0169409. doi:10.1371/journal.pone.0169409.

Davies P. T. and Tso M. S. (1982) Procedures for Reduced-rank Regression. Appl. Statist. 31: 244-255.

Legendre P., Oksanen J. and Ter Braak C. J. F. (2011) Testing the significance of canonical axes in redundancy analysis. Methods in Ecology and Evolution, 2: 269–277.

Naranjo S. E. (2005) Long-Term Assessment of the Effects of Transgenic *Bt* Cotton on the Abundance of Nontarget Arthropod Natural Enemies. Environ. Entomol. 34(5): 1193-1210.

O’Hara R. B., Kotze D. J. (2010) Do not log-transform count data. Methods in Ecology and Evolution, 1: 118–122.

Paliy O. and Shankar V. (2016) Application of multivariate statistical techniques in microbial ecology. Molecular Ecology, 25: 1032–1057.

SAS (2012). Software Release 9.4 (TS1M4). Cary, North Carolina, Copyright 2002-2012 by SAS Institute, Inc.

St-Pierre A. P., Shikon V., Schneider D. C. (2018) Count data in biology-Data transformation or model reformation? Ecology and Evolution, 8: 3077-3085.

Smilauer P., Leps J. (2014) Multivariate Analysis of Ecological Data Using CANOCO 5. Cambridge University Press, Cambridge.

Ter Braak C. J. F. and Looman C. W. N. (1994) Biplots in Reduced-Rank Regression. Biom. J. 36(8): 983-1003.

Ter Braak C. J. F. and Smilauer P. (1998a) CANOCO reference manual and user’s guide to CANOCO for Windows: software for canonical community ordination (version 4). Microcomputer Power, Ithaca, NY.

Ter Braak C. J. F. and Smilauer P. (1998b) Multivariate analysis of stress in experimental ecosystems by Principal Response Curves and similarity analysis. Aquatic Ecology, 32: 163–178.

Ter Braak C .J. F. and Smilauer P. (2015) Topics in constrained and unconstrained ordination. Plant Ecology, 216: 683–696 .

Van den Brink P. J. and Ter Braak C. J. F. (1998) Multivariate analysis of stress in experimental ecosystems by Principal Response Curves and similarity analysis. Aquatic Ecology, 32: 163–178.

Van den Brink, P. J., and Ter Braak C. J. F. (1999) Principal response curves: analysis of time-dependent multivariate responses of biological community to stress. Environmental Toxicology and Chemistry, 18: 138-148.

Vendrig N. J., Hemerik L., Ter Braak C. J. F. (2017) Response variable selection in principal response curves using permutation testing. Aquatic Ecology, 51: 131-143.

Warton D. I. (2005) Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. Environmetrics, 16(3): 275-289. Doi:10.1002/env.702

Fig 1. Graphical representations of sample mean difference (Diff) and standard error (SE) and t ratio (= Diff/SE) as functions of the expected mean count with and without transformation. Two random samples of $n = 10$ were generated from two negative binomial distributions with mean differing by 25% and average of two means ranging $(0.1{\sim}10)$ and variance of each distribution $\sigma^2 = \mu + a\mu^2$ as $a = (0, 0.1, 0.2)$. Scales include count (Count), square root (SQRT), and logarithm (Log). Each point represents a mean of 10000 repeats.

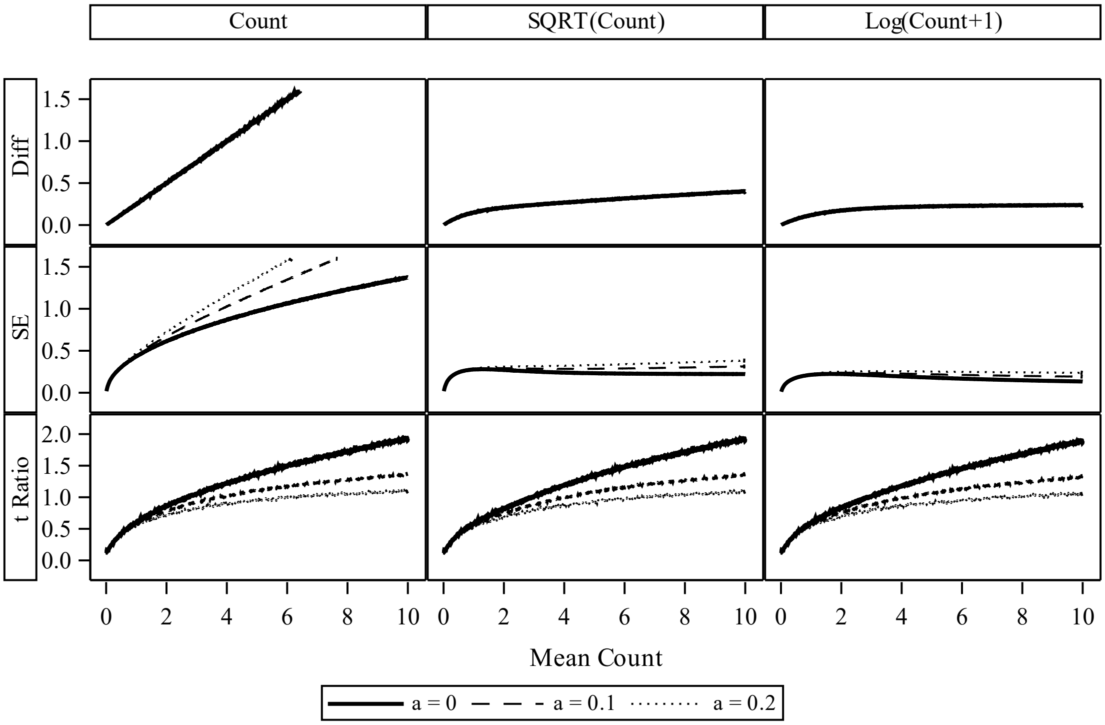
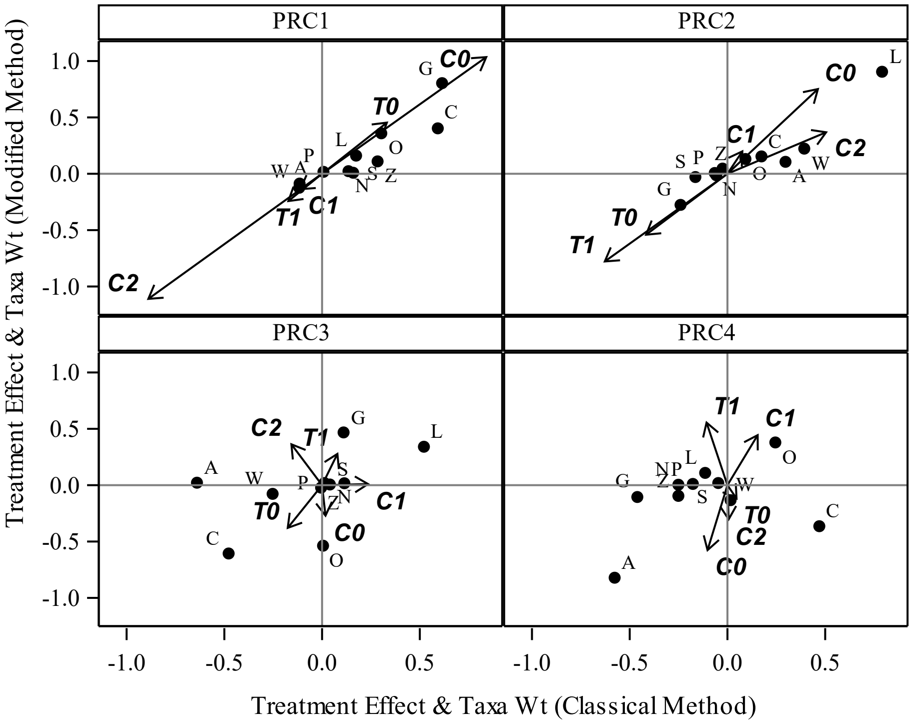


Fig 2. Ordination diagrams of PRC analyses of MON 88702 data: Comparing the estimated main treatment effect (arrow)and taxa weight (dot) of PRC1-4 between the classical and modified methods. Labels were described in text for the treatment, and by initials of Table 1 for the taxa. To fit taxa weight b_k and the treatment effect c_{ilm}^{a} in the same plot, the treatment effect was multiplied by 0.75, and rescaled for PRC2-4 in reciprocally proportional to the square root of the variation relative to that of PRC1.


PRC1
PRC2
PRC3
PRC4
Treatment Effect & Taxa Wt (Modified Method)
G C0
C
T0
L
O
P
W A
S Z
N
T1 C1
C2
L
C0
C1
S P Z
C
C2
W
O
A
N
G
T0
T1
G
C2
T1
L
A
W
P
S
Z N
C1
T0
C0
C
O
T1
C1
NP L
G Z
S
W
O
T0
C2
C
C0
A
Treatment Effect & Taxa Wt (Classical Method)