Dimension Reduction for Multinomial Models Via a Kolmogorov-Smirnov Measure (KSM)

Stephen C. Loftus, Leanna L. House, Myra C. Hughey, Jenifer B. Walke, Matthew H. Becker, Lisa K. Belden

Abstract

Due to advances in technology and data collection techniques, the number of measurements often exceeds the number of samples in ecological datasets. As such, standard models that attempt to assess the relationship between variables and a response are inapplicable and require a reduction in the number of dimensions to be estimable. Several filtering methods exist to accomplish this, including Indicator Species Analyses and Sure Information Screening, but these techniques often have questionable asymptotic properties or are not readily applicable to data with multinomial responses. As such, we propose and validate a new metric called the Kolmogorov-Smirnov Measure (KSM) to be used for filtering variables. In the paper, we develop the KSM, investigate its asymptotic properties, and compare it to group equalized Indicator Species Values through simulation studies and application to a well-known biological dataset.

1 Introduction

New technology has provided opportunities to collect different and more data than ever before. For example, technological advancements in mircroarrays and spectrometry have enabled the collection of hundreds or thousands of molecular measurements in a single sample. Alas, developments in methodology for analyzing large datasets have not kept up with the technology. Many standard analytical approaches breakdown due to computational or algorithmic constraints. A common issue is that modern datasets include more dimensions, i.e., variables, than observations. Hence, models that aim to assess the relationship between the variables and a response are inestimable; the models are oversaturated and standard estimators of model parameters are not unique. To reduce dimensionality and enable modeling, we propose and validate a new metric, called the Kolmogorov-Smirnov Measure (KSM) that may be used for filtering variables.

The KSM is useful when attempting to assess the relationship between P variables and a multinomial response y. It allows analysts to pre-screen variables individually and select those that have the most potential of explaining significant variation in y. By pre-screening, analysts reduce the number of variables and thereby stabilize parameter estimates in subsequent models of y, e.g., regression models. Broadly, the KSM adjusts the calculation of a common statistic, known as the Kolmogorov-Smirnov Statistic (Kolmogorov, 1933; Smirnov, 1936), to summarize differences in empirical distributions. When a response y may take K distinct values, there are K empirical distributions per variable. The KSM quantifies the degree to which the K empirical distributions differ. Variables with a high KSM are kept for further analysis, whereas variables with a low KSM are dropped.

There are several current methods for filtering variables. However, many such methods assume the variables are Gaussian and rely on K measures of central tendency per variable to assess its potential importance. By using the KSM, we relax the assumption of normality in that the KSM accounts simultaneously for distributional shape, outliers, central tendency, and variance. Also, many filtering methods do not apply easily to multinomial responses, including Sure Information Screening (Fan and Lv, 2008), and Indicator Species Analysis (De Cáceres et al., 2010). Yet, the KSM may apply when the response has $K \geq 2$ values in its domain.

The remainder of the paper proceeds as follows. In Section 2, we motivate our research, briefly describe a common filtering approach called the group-equalized Indicator Species Analysis (De Cáceres et al., 2012), and develop KSM. Because the group-equalized Indicator Species Analysis is widely accepted in the ecology community, it provides a natural bench mark to which we compare asymptotic properties of KSM and the application of KSM in simulation and real-data scenarios. We present the results of the applications in Section 3. Finally, we conclude with a discussion of the results and implications of the KSM in Section 4.

2 Methods

Our work was motivated by both a taxonomic application and limitations presented by common analytical methods for such data. In this section, we use the application to establish some notation and vocabulary, as well as, provide background information for common filtering methods. In particular, we explain the group-equalized Indicator Species Values (De Cáceres et al., 2012), a popular approach in ecological studies (Zhang et al., 2014; Suz et al., 2014, e.g.,), so that we may ultimately compare the approach to the KSM.

Taxonomic datasets contain measurements for N samples that reflect the abundance of P operational taxonomic units (OTUs). An OTU is a set of barcoded DNA sequences that are identical by a user-defined percent or more; e.g., one OTU may include several DNA sequences that are at least 97% the same. In our work, we use OTU measurements as a reflection of the microbial communities that live on the skin of amphibians. For a preliminary analysis, we want to understand differences in communities on frog species.

For typical datasets, when the number of variables P is less than the number of observations N, we could fit a generalized linear model to quantity the relationship between OTUs and species. That is, if we let y_i represent the species of frog i and x_i represent an OTU vector of measurements, ($i \in \{1, 2, ..., N = 137\}$), we might fit a multinomial probit model to assess which OTUs are significant, e.g.,

$$\begin{aligned} y_i &\sim \text{Multinomial}(\rho_1, \rho_2, ..., \rho_K) \\ \Phi^{-1}(p_k) &= \beta_0 + x_i\beta, \end{aligned} \tag{1}$$

where, $\Phi^{-1}()$ represents a standardized Normal inverse cumulative distribution function; p_k ($k \in \{1,,,,K\}$) represents the probability $y_i = k$; β_0 presents an intercept term; and β represents an $P \times 1$ vector of variable coefficients. Estimates of the coefficients reflect the relationships between each variable and y. Unfortunately, in many taxonomic datasets, the number of variables exceeds the number of observations by several orders of magnitude; $P >> N$. For such datasets, Model 1 is oversaturated and parameter estimates are not unique. To resolve the problem, we may either constrain model parameters in the estimation process or apply a filtering method to reduce the number of variables to consider. Both options have pros and cons. In this paper, we opt to develop KSM: an easy-to-implement, easy-to-interpret, and effective means to filter variables. For a means of comparison, we first summarize group-equalized Indicator Species Analyses (geISA).

2.1 Group-Equalized Indicator Species Analysis

Group-equalized Indicator Species Analyses (geISA) De Cáceres et al. (2012) is a common method for filtering variables that emerged from previous versions (Dufrêsne and Legendre, 1997; De Cáceres et al., 2010). The name stems from its motivating application to identify species that appear consistently in one of any number of response groups, but the approach may apply in a variety of settings and is available in the statistical software R, via the *indicspecies* package (De Cáceres and Legendre, 2009). Similar to KSM, geISA relies on measuring the association of each variable with the response and selecting those with the largest association. The measure is called the Indicator Species Value (ISV) and, in the context of taxonomic data, it is calculated as follows.

Let Y represent an $N \times 1$ vector of responses such that each $y_i \in \{1, ..., K\}$ ($i \in \{1,,,.N\}$) and let X represent a community data matrix X with N observations and P variables, i.e., OTUs. The ISV of a variable j ($j \in \{1, ..., P\}$) at a response level k is the product of a positive predictive value

A and measure of sensitivity B,

$$\text{ISV}_{j,k} = A_{j,k}B_{j,k}.$$

The positive predictive value $A_{j,k}$ reflects the proportion of OTU abundance measured per response level and variable. It is calculated by dividing the average abundance for variable j at response level k, $\bar{x}_{j,k}$, by the sum of average abundances in all groups,

$$A_{j,k} = \frac{\bar{x}_{j,k}}{\sum_{l=1}^{K} \bar{x}_{j,l}}. \tag{2}$$

Note, this positive predictive value deviates slightly from previous, nonequalized ISV versions (Dufrêsne and Legendre, 1997; De Cáceres et al., 2010) in that it uses group averages rather than groups sums. Also, $A_{j,k}$ can be calculated for presence-absence or abundance data, although the abundance data holds more information and is recommended when possible. The sensitivity measure $B_{j,k}$, is the proportion of non-zero abundance values at response level k for variable j,

$$B_{j,k} = \frac{n_{j,k}}{N_k}, \tag{3}$$

where $n_{j,k}$ is the number of non-zero abundance values of variable j for response level k and N_k the number of observations for response level k.

To filter the variables based on ISV calculations, there are two primary options. The first uses simulation to develop an ISV distribution for each variable and calculate a p-value for the observed ISV. Those deemed significant (e.g., those variables with a p-value less than 0.05) are kept for further analysis. However, such an approach is computationally intensive and may result in too many or too few variables from the dataset. The second method is based of user-defined thresholds for both quantities $A_{j,k}$ and $B_{j,k}$. If either quantity falls above the assigned thresholds, the variable is kept for further analysis. Otherwise, the variables removed. The specification for thresholds may rely on expert judgement, computational constraints, and/or sensitivity analyses.

In standard cases, the ISV has shown effective, but it has peculiar asymptotic properties that can be misleading. In practice, the ISV has a continuous interpretation in that large ISVs suggest large differences in variable distributions across response levels, relative to variables with small ISVs. However, there is a constraint on this interpretation of the ISV. When distributions are equal, the ISV approaches $1/K$, not zero, as the number of observations in each response level, N_k, increases. Namely, as each N_k goes to infinity, the ISV becomes a function of expectations and probabilities, whereby

$$\lim_{N_1,\ldots,N_K \to \infty} \text{ISV}_{j,k} = \frac{E(x_j|y_k)}{\sum_{i=1}^{K} E(x_j|y=i)} \times P(x_j \neq 0|y=k) \tag{4}$$

(Supplemental Information S1). Under the assumption that a variable is sufficiently bounded away from 0, $P(X_j \neq 0|y=k) \;\to\; 1$ and, when the variable is independent of response level we have

$$E(x_j|y=k) = E(x_j|y=k') \;\forall\; k \;\neq\; k'.$$

Thus, $\text{ISV}_{j,k}$ in Equation 4 goes to $1/K$ as $N_k \;\to\; \infty \,\forall\, k$. Converging to $1/K$ rather than zero poses a problem because, for some datasets, the value of $1/K$ may be large enough to warrant including in subsequent analyses; methods based on ISVs are vulnerable to high false positive rates.

Further, Equation 4 shows that ISVs depend entirely on measures of central tendency, specifically the expectation (which is estimated by a sample mean). In many instances, sample means may not describe differences between two distributions sufficiently. For example, consider a response variable with two categorical levels, $y \in \{1, 2\}$ and two variables x_1 and x_2. Let x_1 and x_2 be independent and dependent of y, respectively, so that $x_1 \sim \text{Gamma}(10, 1)$, $x_2|y = 1 \sim \text{Gamma}(100, 10)$, and $x_2|y = 2 \sim \text{Gamma}(1, 0.1)$. The expectations of each distribution is the same, $E(x_1|y = 1) = E(x_1|y = 2) = E(x_2|y = 1) = E(x_2 y = 2) = 10$, and both variables have the same ISV for each response level, $\text{ISV}_{1,1}$=$\text{ISV}_{1,2}$=$\text{ISV}_{2,1}$=$\text{ISV}_{2,2}$=0.5. However, clearly, the distributions generating x_2 are vastly different, as shown in Figure 1. Also, if we were to use the ISV to detect x_2 as dependent on y, we would also declare x_1 dependent on y.

The intent for the ISV measure is to reflect the degree to which distributions are similar and different, and it would make sense for the measure to improve as more data are collected. Yet,

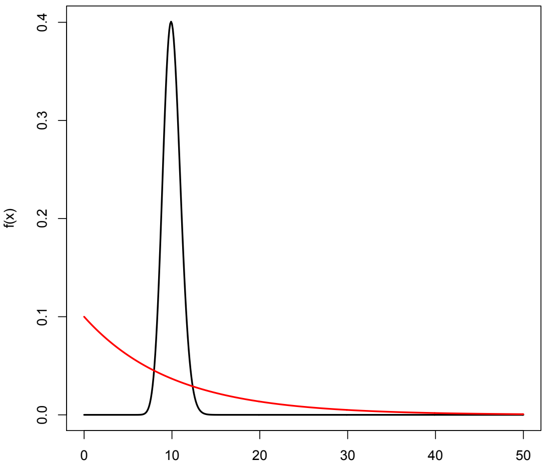

Figure 1: Two different Γ distributions with identical expectations

because of its heavy reliance on measures of central tendency per distribution, the ISV may be misleading even when sample size increases. In the next subsection we describe our approach in detail to assesses entire, empirical cumulative distribution functions and measure distributional differences.

2.2 Kolmogorov-Smirnov Measure (KSM)

Our Kolmogorov-Smirnov Measure (KSM) is an extension of an early test statistic that is used to compare two distributions, the Kolmogorov-Smirnov (KS) test of Kolmogorov (1933) and Smirnov (1936). To develop the KSM, we first describe the KS test.

Instead of looking for differences in sample averages or medians, the KS test looks for the largest difference in the empirical distributions $\hat{F}$ of two samples. Given one response $y \in \{1, 2\}$ and one variable x_j the KS test statistic of the KS test is defined as

$$\begin{aligned} KS_j &= \max_a |\hat{F}_{j,1}(a) - \hat{F}_{j,2}(a)| \qquad (5) \\ &= \max_a \left| \frac{\sum_{i=1}^{N_1} \mathbb{1}\{x_{j,1,i} \leq a\}}{N_1} - \frac{\sum_{i=1}^{N_2} \mathbb{1}\{x_{j,2,i} \leq a\}}{N_2} \right| \end{aligned}$$

where $\hat{F}_{j,k}$ is the empirical distribution function of x_j when $y = k$, N_k is the number of observations at response level k, and $x_{j,,k,i}$ is the i^{th} sample of x_j when the response level equals k. As the KS test is often phrased as a typical hypothesis test, significant differences between distributions are often declared by comparing a p-value to a type I error rate.

In the context of our taxonomic datasets, KS test may only apply when $K = 2$; i.e., there are only two distributions to compare. Yet, the response in our datasets includes more levels to consider, i.e., $K > 2$, so we cannot use the KS test as it stands to identify important variables in explaining a multinomial response y . However, there are two features of the KS test that we may exploit extend it to our applications. One, the KS test statistic conforms to the definition of a distance metric and, two, because empirical distribution functions are bounded between [0,1], the KS test statistic is also bounded between [0,1]. Values close to one indicate a more pronounced difference in distributions between samples than values close to zero. With these features in mind, we define the KSM to be a weighted sum of all $\binom{K}{2}$ pairwise KS test statistics per variable. The weights are chosen to be proportional to the total sample size in the comparison, so that the final KSM for one variable equals

$$KSM_j = \sum_{k=1}^{K} \sum_{k \neq k'} \frac{N_k + N_{k'}}{N \times (K-1)} KS_j(k, k'). \qquad (6)$$

where $KS_j(k, k')$ is the KS test statistic for comparing the distributions of one variable x_j when the response equals k or k' and N is the total sample size for the study. Because the weights sum to one, the KSM is also bounded between [0,1] and has the same interpretation as the KS test statistic; large value close to one, reflect large differences between the distributions of x_j when the response takes multiple levels.

There are a variety of approaches to filter variables based on KSMs. In general, we recommend imposing a user-defined threshold so that all variables with a KSM greater than the threshold are kept for further analysis and those less than the threshold are removed. When possible, we also recommend sensitivity analyses in the threshold choice. Our preferred approach in selecting the threshold begins by plotting each KSM in ranked order and identifying drop points in the curve that suggest a sudden change in the degree to which variables relate to the response. In turn, we set the threshold for the KSM at the drop point. Refer to Sections 3.2 and 3.3 for an example. Often, KSM plots will present several drop points (i.e., several suggestions for KSM thresholds). To decide among them, we recommend sensitivity analyses that are relevant to the subsequent analyses of the data; e.g., changes or lack there of in data visualizations, model cross-validation results, and/or measures of model fit, e.g., the Bayesian Information Criterion (Schwartz, 1978) may assist in the selection of the KSM threshold. Also, it is important to consider computational constraints when selecting the threshold. Uses could define the threshold to include the maximum number of variables that will still result in stable parmeters estimates in subsequent analyses. In turn, subsequent analyses may remove variables; e.g., tests of significance in generalized linear models.

The KSM is easy interpret and apply for any dataset with multinomial responses. In fact, we created an `R` package called *KSmeasure* that is available for download here: `http://www.apps.stat.vt.edu/house/research.html`. Additionally, the KSM has asymptotic properties that reinforce its utility, even as sample size N increase. Namely, unlike the ISV, when a variable x is independent of the response y, the KSM goes to zero and nonzero otherwise. To see this, we know by the Glivenko-Cantelli Theorem (Glivenko, 1933; Cantelli, 1933) that the KS test, $KS_j(k, k')$ for one variable x_j and two-level response equals

$$\lim_{N_k, N_{k'} \to \infty} KS_j(k, k') = \max_a |F_{j,k}(a) - F_{j,k'}(a)|. \tag{7}$$

Thus, the limit of the KSM equals

$$\lim_{N_1 \ldots N_K \to \infty} KSM_j = \frac{1}{\binom{K}{2}} \sum_{k=1}^{K} \sum_{k \neq k^prime} \max_a |F_{j,k}(a) - F_{j,k^prime}(a)|. \tag{8}$$

Notice that this limit equals zero if and only if $f(x_j|y=1) = ... = f(x_j|y=K) = f(x_j)$. In other words, if there are differences in distribution of x_j given the response level of y, the KSM will be asymptotically non-zero, whereas as if the distribution of x_j is independent of response level then $KSM \to 0$ with probability 1. We show this result in practice via simulation studies in the next section.

3 Results

To demonstrate the use of the KSM and compare it to the ISV, we present analytical results for both methods in this section. We use simulation studies and real-world applications to highlight the theoretical and practical differences between the two methods.

3.1 Simulation Studies

Two simulations are designed to illustrate two asymptotic properties of the ISV and KSM. The first simulation illustrates the constant non-zero value for the Indicator Species analysis for identical distributions. The second simulation illustrates the reliance of Indicator Species values solely on measures of central tendency, regardless of the similarity of distributions.

For the first simulation, a response vector with three categorical levels is created; $y \in \{1, 2, 3\}$. In this simulation, the distribution of $f(x_j)$ is independent of response level so that $f(x_j) = \Gamma(1000, 10)$. Samples were drawn from this distribution, with $N_1 = N_2 = N_3$ and ranging in value from 10 to 10,000. As seen in Figure 2(a), the geISA remains at $\frac{1}{3}$, whereas the KSM goes to zero.

In the second simulation, a two-level categorical response ($y \in \{1, 2\}$) is created, but the distribution of x_j is dependent on response level. In this case, $f(x_j|y=1) = \Gamma(1000, 10)$ while $f(x_j|y=1) = \Gamma(100, 1)$, so that $E(x_j|y=1) = E(x_j|y=2) = 100$, but $Var(x_j|y=1) = 10$ and $Var(x_j|y=2) = 100$. Then, samples of 100 observations each were drawn from each distribution

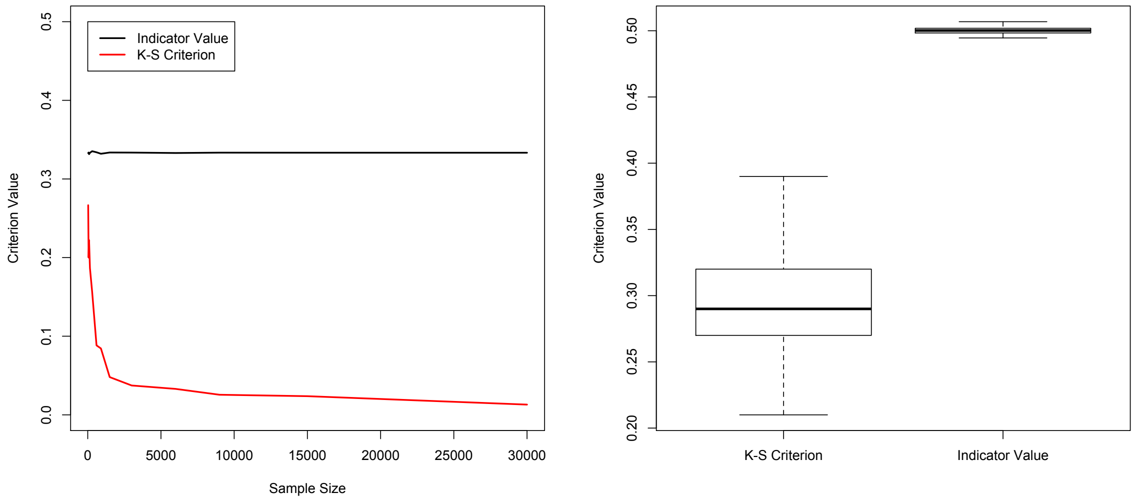

Figure 2: Simulation results. (a) Asymptotic values of ISV and KSM as $n \to \infty$. (b) ISV and KSM values for two different distributions that have identical expectations

($N_1 = N_2 = 100$) 10000 times. Figure 2(b) presents boxplots of the simulated geISA and KSM estimates. Theoretically, the geISA and KSM should be around 0.5 and 0.2623, which can be calculated through equation 4 and equation 7. Both geISA and KSM approach these values, as geISA centers around $\frac{1}{2}$ and ranges between $[0.48, 0.52]$ whereas, the KSM centers around 0.28 and ranges between [0.21, 0.39]. This presents a difficulty for geISA as the value for two different distributions is approximately the value for two differing distributions.

The KSM has a clear advantages in simulation. Next, we apply the KSM to a common dataset used to explain the ISV in the literature.

3.2 Oribatid Mite Dataset

Consider a common dataset, know as the Oribatid mite dataset that was collected by Daniel Borcard in 1989 (Borcard et al., 1992). This dataset is available through the *vegan* package in R (Oksanen et al., 2015). In this instance, the community data matrix includes counts on 35 species ($P = 35$) of Oribatid mites on 70 observations ($N = 70$), and a categorical response describing the microtopography of the observation, either Blanket of Hummock ($K = 2$). We calculate the ISV and KSM for each species and plot them in ranked order in Figure 3.

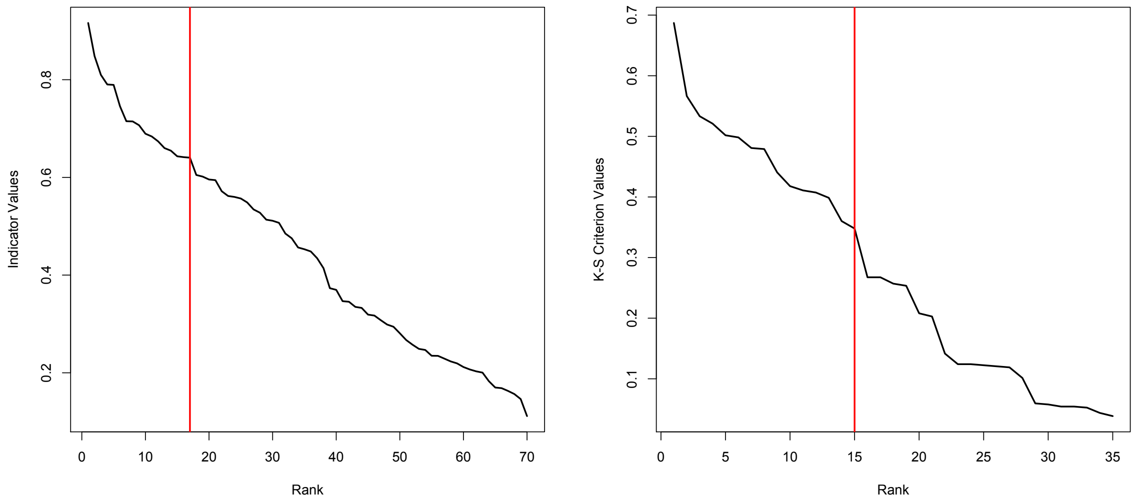

Figure 3: Cutoffs for Indicator Values (On Left) and K-S Measure (On Right)

For both methods, we identify a natural drop in the plots to determine thresholds for filtering variables. For the ISV, we see a drop at the 17th ranked species, thus all species with $ISV > 0.62$ are kept for further analysis. For the KSM, we see a drop at the 15th ranked species, thus all species with $KSM > 0.35$ are kept for further analysis. In the two sets of selected variables, 12 overlap. For each set, we fit a probit model to assess the selected variables further and in relation to the response. Based on tests of significance with a generous type I error ($\alpha = 0.1$), three of the variables selected by the ISV and five of those selected by the KSM are deemed significant. Additionally, we refit the probit models without 20 observations to assess prediction accuracy. The variables selected by the ISV predicted 65% (13 out of 20) accurately, while the KSM predicted 85% (17 of 20) accurately.

From both a theoretical and practical perspective, the KSM shows advantages over the ISV. Thus, in the next subsection, we complete our final application using just the KSM.

3.3 Panamanian Tree Frog Dataset

In this dataset (Belden et al., 2015)., we have data from, $N = 137$ Panamanian frogs of three species that were captured and swabbed within 30 minutes. Of the 137, 62 were *Agalychnis callidryas*, 22 were *Craugastor Fitzingeri*, and 53 were *Dendropsophus ebraccatus* that we reference as AgaCal, CraFit, and DenEbr, respectively. The frogs were collected from four sites—Parque Nacional Altos de Campana, Parque Nacional Soberanía, Mamoní Valley Preserve, and the forest surrounding Nuevo Vigía—with three species represented—*Agalychnis callidryas* (Also notated AgaCal), *Craugastor fitzingeri* (CraFit), and *Dendropsophus ebraccatus* (DenEbr). Although almost all the frogs had been exposed to Bd at some time, only 36 of the frogs were infected with Bd as indicated by low measures of Bd zoospores.

The frogs were captured by hand and the swabbed within 30 minutes of capture. Capture locations were dependent on the species but conformed to certain specifications. For example, *Agalychnia callidryas* and *Dendropsophus ebraccatus* are both pond breeders, and these species were sampled from a single pond at all four sites. *Craugastor fitzingeri* individuals were sampled on the margins of a single stream at the two sites where the species was encountered.

Prior to swabbing, the frogs were rinsed with sterile deionized water. Each swab consisted of 40 strokes over various, pre-set locations on the frog. Upon returning to the United States, the swabs were analyzed by the program Quantitative Insights Into Microbial Ecology (QIIME) and rarified to a depth of 7,000 total reads Heck et al. (1975). In turn, DNA sequences that were at least 97% the same were grouped, summed, and scaled between zero and one. The end result was a dataset containing 3,490 distinct OTU measurements per frog that were transformed to a [0,1] scale by calculating relative abundance measurements.

The KSM was calculated for each of the 3,490 OTUs and plotted in ranked order in Figure 4. We see a natural cutoff at the 42nd ranked OTU (threshold=0.41). Although, we see potential thresholds at the 23rd and 27th ranked OTUs, we opt to apply multinomial logistic regression to filter variables further from a set of 42. Using t-tests of significance for model coefficients with $\alpha = 0.05$, 35 OTUs see important in explaining species. To assess the predictive ability of the 35 OTUs, a holdout sample of 37 frogs was selected, and a multinomial training model was fit for the

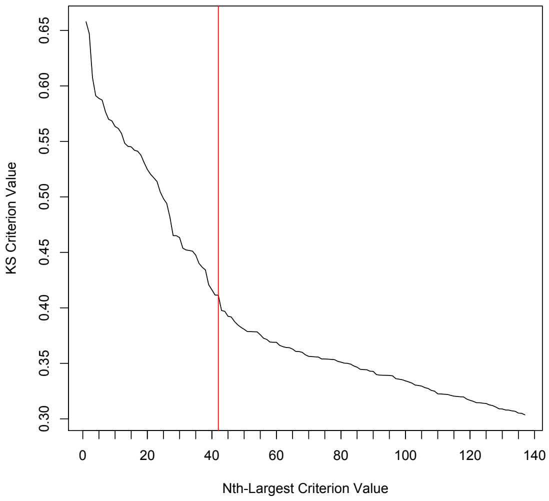

Figure 4: Top KSMs for the Panamanian Tree Frog Dataset. A “natural cutoff” is plotted by the vertical line.

remaining 100 frogs. The species holdout sample was predicted based on the coefficients of the training model, and the prediction accuracy was calculated. The 35 OTUs had a 89.2% prediction accuracy, with the specific prediction results given in Table 1. We miss classify one AgaCal and three DenEbr.

For further validation we visualize the initial and reduced datasets nonmetric multidimensional scaling (NMDS) (Kruskal, 1964a,b). NMDS with a Bray-Curtis dissimilarity measure has shown useful for datasets with properties presented by Panamanian Tree Frog Dataset (Minchin, 1987). To create the NMDS plots, we apply the *metaMDS* function from the *vegan* package in `R`. The initial dataset resulted in Figure 5(a) with a final stress of 0.1159, while the cut dataset resulted in Figure 5(b) with a final stress of 0.0587. The structure seen in the initial NMDS is retained in the cut dataset. AgaCal overlaps with both CraFit and DenEbr, although mostly with DenEbr, while CraFit and DenEbr do not overlap whatsoever. This implies that the cut dataset retains a majority of the information found in the complete dataset; the KSM is able to accurately select the variables important for modeling a response.

Table 1: Cross-Validation of Taxonomic Data Model

		True Species		
		AgaCal	CraFit	DenEbr
Predicted Species	AgaCal	13	0	1
	CraFit	0	5	0
	DenEbr	3	0	15

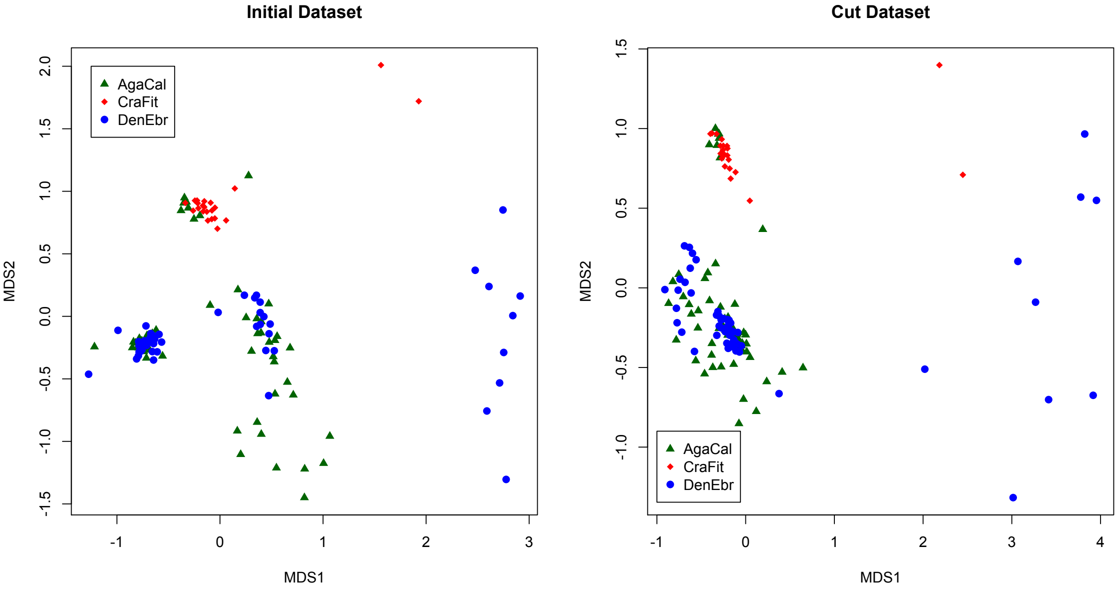

Figure 5: NMDS results for the initial Panamanian tree frog dataset (On the left) and the cut dataset (On the right)

4 Discussion

In this paper, we introduced the Kolmogorov-Smirnov Measure (KSM) as a means to filter variables in large datasets for further investigation of a multinomial response. The KSM is easy to apply and interpret and is available in an R package (Loftus, 2015). We justify the use of the KSM from both theoretical and practical perspectives, as well as compare it to a common filtering approach in the Ecology applications called group-equalized Indicator Species Analyses (geISA). For brevity, we did not discuss other filtering and dimension reduction techniques in detail, including multiple t-testing, Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996), Sure Information

Screening (Fan and Lv, 2008), and principal component analysis (Pearson, 1901; Hotelling, 1933) . Such methods require distributional assumptions of the variables, fail to apply well in multinomial settings, and/or can be hard to interpret.

In geISAs, a metric known as an Indicator Species Value (ISV) is used in a similar way as the KSM to filter variables. Variables with a high ISV or KSM are considered strongly associated with a response and kept for further analysis. However, we showed that properties of the ISV may lead to higher false positive rates than the KSM and, in some applications, the KSM may result in higher predictive rates than the ISV. Our primary application for the KSM was to analyze the Panamanian tree frog dataset, with the goal of attempting to discern OTUs that distinguish between three species of tree frogs. After cutting the dataset down from nearly 3,500 OTUs to a subset of 35, the resulting OTUs accurately predicted 89.2% of a holdout sample of 37 frogs. This finding was further supported by nonmetallic multidimensional scaling (NMDS) plots for both the full and cut Panamanian tree frog datasets. Effectively, little information was lost in reducing the datasets by the KSM.

In our applications, we were interested in variables that associated strongly with a response; i.e., variables that presented high KSMs. However, the KSM can also be used to identify variables that are highly similar across response levels. For example, in the Panamanian tree frog dataset, it may be worth investigating similarities and differences in the OTUs that did not explain variation in species.

To conclude, we admit that the KSM suffers from one drawback that is shared by many filtering methods. The KSM investigates variables individually, rather than jointly. Thus, ideal conditions for the KSM is that all variables are independent and the variables work independently to explain variation in a response. Yet, such conditions are rarely met in modern high-dimensional datasets. The implication of the drawback is that the KSM can be inefficient in selecting a parsimonious set of variables to explain variation in a response; any variable associated with a response, regardless of its association with others, will have a high KSM. A worst case scenario is that the KSM keeps two or more identical variables for further analysis because they each associate strongly with the response. This may or may not pose a problem in subsequent analyses and practitioners should be aware. There are some ways to weaken the implications, such as create interaction variables and apply the KSM on them. We leave such developments for further research. The simplicity and effectiveness of the KSM still support its use in exploratory or pre-screening analyses.

References

Belden, L., Hughey, M., Rebollar, E., Umile, T., Loftus, S., Burzynski, E., Minbiole, K., House, L., Jensen, R., Becker, M., Walke, J., Medina, D., Ibáñez, R., and Harris, R. (2015), "Structure-function relationships in the bacterial skin microbiome of Panamanian amphibians," Work in progress.

Borcard, D., Legendre, P., and Drapeau, P. (1992), "Partialling out the spatial component of ecological variation," *Ecology*, 73, 1045–1055.

Cantelli, F. (1933), "Sulla determinazione empirica della legge di probabilita," *Giorn. Ist. Ital. Attuari*, 4, 421–424.

De Cáceres, M. and Legendre, P. (2009), *Associations between species and groups of sites: indices and statistical inference.*

De Cáceres, M., Legendre, P., and Moretti, M. (2010), "Improving indicator species analysis by combining groups of sites," *Oikos*, 119, 1674–1684.

De Cáceres, M., Legendre, P., Wiser, S., and Brotons, L. (2012), "Using species combinations in indicator value analyses," *Methods Ecol. Evol.*, 3, 973–982.

Dufrêsne, M. and Legendre, P. (1997), "Species Assemblages and Indicator Species: The Need for a Flexible Asymmetrical Approach," *Ecol. Monogr.*, 67, 345–366.

Fan, J. and Lv, J. (2008), "Sure independence screening for ultrahigh dimensional feature space," *J. R. Statist. Soc. B*, 70, 849–911.

Glivenko, V. (1933), "Sulla determinazione empirica della legge di probabilita," *Giorn. Ist. Ital. Attuari*, 4, 92–99.

Heck, K., van Belle, G., and Simberloff, D. (1975), "Explicit Calculation of the Rarefaction Diversity Measurement and the Determination of Sufficient Sample Size," *Ecology*, 56, 1459–1461.

Hotelling, H. (1933), "Analysis of a complex of statistical variables into principal components," *J. Educ. Psychol.*, 24, 498–520.

Kolmogorov, A. (1933), "Sulla Determinazione Empirica di una Legge di Distribuzione," *Giorn. Ist. Ital. Attuari*, 4, 83–91.

Kruskal, J. (1964a), "Multidimensional scaling by optimizing goodness of fit to a nonmetallic hypothesis," *Psychometrika*, 29, 1–25.

— (1964b), "Nonmetric multidimensional scaling: A numerical method," *Psychometrika*, 29, 115–129.

Loftus, S. (2015), *Kolmogorov-Smirnov Criterion*, r package version 1.0.

Minchin, P. (1987), "An evaluation of the relative robustness of techniques for ecological ordination," *Vegetatio*, 69, 89–107.

Oksanen, J., Blanchet, F., Kindt, R., Legendre, P., Minchin, P., O'Hara, R., Simpson, G., Solymos, P., Stevens, M., and Wagner, H. (2015), *vegan: Community Ecology Package*, r package version 2.2-1.

Pearson, K. (1901), "On lines and planes of closest fit to systems of points in space," *Philos. Mag.*, 2, 559–572.

Schwartz, G. (1978), "Estimating the Dimension of a Model," *Ann. Statist.*, 6, 461–464.

Smirnov, N. (1936), "Sur la distribution de ω^2 (criterium de M. R. v. Mises)," *C. R. Acad. Sci.*, 202, 449–452.

Suz, L., Barsoum, N., Benham, S., Dietrich, H., Fetzer, K., Fischer, R., Garía, P., Gehrman, J., Krist ofel, F., Manninger, M., Neagu, S., Nicholas, M., Oldenburger, J., Raspe, S., Sánchez, G., Schr ock, H., Schbert, A., Verheyen, K., Verstraeten, A., and Bidartondo, M. (2014), "Environmental drivers of ectomycorrhizal communities in Europe's temperate oak forests," *Mol. Ecol.*, 23, 5628–5644.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *J. R. Statist. Soc. B*, 58, 267–288.

Zhang, L., Jia, S., Yang, M., Xu, Y., Li, C., Sun, J., Huang, Y., Lan, X., Lei, C., Zhou, Y., Zhang, C., Zhao, X., and Chen, H. (2014), "Detection of copy number variations and their effects in Chinese bulls," *BMC Genomics*, 15, 449–452.