STATISTICAL ISSUES IN THE ANALYSIS OF MICROBIAL COMMUNITIES IN SOIL

J. D. Wilbur

C. H. Nakatsu

S. M. Brouder

J. K. Ghosh

R. W. Doerge

See next page for additional authors

Author Information

J. D. Wilbur, C. H. Nakatsu, S. M. Brouder, J. K. Ghosh, and R. W. Doerge

STATISTICAL ISSUES IN THE ANALYSIS OF MICROBIAL COMMUNITIES IN SOIL

J. D. Wilbur[1,2], C. H. Nakatsu[3], S. M. Brouder[3], J. K. Ghosh[1,4], R. W. Doerge[1,2,3]

[1] Department of Statistics, Purdue University, West Lafayette, IN 47907-1399
[2] Computational Genomics, Purdue University, West Lafayette, IN 47907
[3] Department of Agronomy, Purdue University, West Lafayette, IN 47907-1150
[4] Indian Statistical Institute, Calcutta, India

Abstract

Corn and soybean production dominates the agricultural systems of the mid-western United States. Studies have found that when a single crop species is grown continually, without the rotation of other crops, yield decline occurs. At present, this phenomenon, remains poorly understood, but there are possible links to microbial community dynamics in the associated rhizosphere soil. In this study, corn plants were grown in disturbed and undisturbed soils with a 24 year history of growth as a monoculture crop or two crops grown in annual rotation. Characteristic profiles of the microbial communities were obtained by denaturing gradient gel electrophoresis of polymerase chain reaction amplified 16S rDNA from soil extracted DNA. This problem is approached as the statistical analysis of high-dimensional multivariate binary data with an emphasis on modeling and variable selection.

1. Introduction

Growing the same crop in a field every year results in lower yields, on the average, than does the practice of rotating crops (e.g., corn and soybean) between fields. Furthermore, long-term analysis of yields suggests a negative, synergistic interaction between the forces controlling the monoculture yield decline and the yield depression associated with corn grown under no-till residue management (West et al., 1996). Recent efforts to identify the mechanisms of monoculture yield decline have shifted the emphasis from unknown abiotic (i.e., non-living) components of the ecosystem to biotic phenomena (i.e., caused by living organisms) mediated by the *microbial community* present in the *rhizosphere soil* (Chiarini et al., 1998). That is to say, the group of several different microbial populations (i.e., species) that live in portion of soil volume in intimate contact with the growing root system. Historically, the technology available to researchers interested in studying microbial communities has been severely limited. Prior to the development of molecular methods for microbial ecology, researchers were limited to inference based on only a small percentage of the microbial populations in a community: those that could be isolated and cultivated in laboratory media. The percentage of all bacteria that can be cultivated in the laboratory has been estimated to be between 0.1 and 10% and as such any inference based on cultivation techniques is likely to be biased. Advances in biotechnology are providing molecular methods to study the general ecology of microbial communities in the environment. In order to fully appreciate the biological and statistical issues associated with these molecular methods, an understanding of both the cellular structure of microorganisms and the laboratory techniques used to produce microbial community fingerprints is necessary.

2. Microorganisms

In any cell, the genetic information is stored in the chromosome(s) in the form of a double stranded macromolecule called deoxyribonucleic acid (DNA). Most microorganisms have only a single chromosome consisting of a, usually circular, DNA molecule which coils tightly to form a compact structure known as the nucleoid. Around the perimeter of all cells there is a cytoplasmic membrane that regulates the flow of materials in and out of the cell. In addition, most microorganisms have a cell wall located just outside the cytoplasmic membrane that is thicker than the membrane and serves as additional protection for the cell. Inside the cell is the cytoplasm, a mixture of substances and structures that carry out the functions of the cell. Among the most abundant cytoplasmic structures in the cell are ribosomes, each one consisting of ribonucleic acid (RNA) molecules and related proteins. The function of RNA in the cell is to transcribe the genetic information present in the DNA and translate it into proteins. This process, called protein synthesis, takes place at the ribosome.

2.1 *Microbial genetics*

Ribosomes are composed of two similar subunits held together by magnesium bonds. The smaller of the two subunits contains the 16S rRNA molecule, which consists of sequences of

nucleotides that are highly conserved among related microorganisms along with a few more variable sequences. This structure allows for effective discrimination between different microorganisms (Pace et al., 1986). For example, some of the more highly conserved nucleotide sequences common to all bacteria can be used to distinguish bacteria from other organisms, while some of the more variable sequences can be used to discriminate between the many different species of bacteria.

Much of the molecular methodology presently used in microbial ecology utilizes the 16S rRNA molecule. Other factors which make the choice of 16S rRNA molecule favorable are the natural amplification of this molecule within the organism due to the large number of ribosomes in each cell, and the availability of the Ribosomal Database Project (RDP-II) (Maidak et al., 2001), which contains over 16,000 different 16S rRNA sequences for comparison. Likewise, the regions of the bacterial chromosome, referred to collectively as the 16S rDNA, which correspond to (or code for) the 16S rRNA, have these same advantageous properties and are often used in microbial community analysis.

The 16S rDNA, like all other DNA molecules are composed of two complimentary strands of nucleotide bases: the forward, or $5' \rightarrow 3'$, strand and the reverse, or $3' \rightarrow 5'$, strand. In DNA, each nucleotide base is made up of 3 components: the sugar, deoxyribose; the phosphate group; and one of four possible nitrogenous bases: Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). The two strands of the DNA molecule are held together by hydrogen bonds between complementary bases, A bonding with T, and C bonding with G. Therefore, if enough energy is applied to a DNA molecule the two strands will separate. This denaturation can be caused by raising either temperature or the pH of the nucleic acid solution or by adding a chemical denaturant such as urea or formamide. Thermal denaturation is often referred to as melting and the temperature at which the two strands will separate is referred to as the T_m of the sequence. The T_m, the pH or the concentration of chemical denaturant at which the two strands of a particular DNA molecule will begin to separate is determined by the length and composition of its nucleotide sequence. In general, more energy is required for separation of longer DNA sequences than for shorter sequences. However, separation of DNA fragments of similar lengths may require more or less energy depending on their nucleotide sequence composition. This is due to the differing number of hydrogen bonds that join nucleotide base pairs. A and T nucleotide bases are joined by 2 hydrogen bonds, while C and G nucleotide bases are joined by 3 hydrogen bonds. Therefore, in general, more energy is required to denature so-called "G+C rich" DNA fragments than for those with a smaller proportion of G and C nucleotide bases.

3. Microbial community fingerprinting

Detailed laboratory protocols for extraction and purification of microbial community DNA from a number of different environmental media (e.g., soil) have been detailed in Akkermans et al. (1996). In general, the process begins with the isolation of the organisms of interest by centrifugation. Then the cells are *lysed*, or broken open, in order to release the chromosomal

DNA from each organism in the solution. The microbial community DNA is then separated from the remainder of the solution by centrifugation.

Having obtained a purified nucleic acid solution from the the environmental sample, polymerase chain reaction (PCR) (Saiki et al., 1985) is applied to the solution in order to amplify the number of copies of some specific region of the 16S rDNA available for analysis. PCR is a cyclical process for DNA replication and in this situation it is used to amplify the number of copies of a specific region of the 16S rDNA that is known to be highly variable between bacterial species.

Denaturing gradient gel electrophoresis (DGGE) can then be applied to the PCR amplified fragments of the microbial community 16S rDNA. This technique, and to a lesser extent thermal gradient gel electrophoresis (TGGE), are now commonly used by researchers in an increasing number of fields to produce characteristic profiles of microbial communities (Muyzer and Smalla, 1998). DGGE is a method by which DNA fragments of similar length can be separated on the basis of their nucleotide sequence composition in a polyacrylamide gel containing a linearly increasing concentration of chemical denaturant. Each fragment, aided by an electric field, migrates down the porous polyacrylamide gel until it reaches the location in the gel at which the concentration of denaturant causes partial separation of the two strands of the DNA fragment. When the two strands begin to separate, the fragment becomes too large to migrate any further down through the polyacrylamide and it stops. TGGE operates on the same principle, but rather than using a gradient of the concentration of chemical denaturant, it uses a thermal gradient to separate the DNA fragments on the basis of their T_m. When the electrophoretic process has been completed the polyacrylamide gel is stained with a fluorescent dye, such as Ethidium bromide or SYBER Green I, that illuminates the characteristic profile of DNA fragments present in the gel.

Since the objective of microbial community DNA fingerprinting is to produce a characteristic profile of the community of interest, we can take advantage of the aforementioned molecular technology to produce the necessary data. Most often this profile takes the form of a lane of illuminated bands which indicate the presence of the different microbial populations in the community. These fingerprints can then be used as a basis for between community comparisons. Figure 1 illustrates an example of representative profiles from rhizosphere soil communities extracted from corn grown under four different agronomic treatments.

In practice, characteristic profiles for approximately 12-20 different communities can be generated side by side in vertical lanes on a single gel. Each gel contains at least one profile of a nucleic acid solution of known composition, and is referred to as the standard or "marker" lane that facilitates comparisons between lanes on different gels. In order to make comparisons between community fingerprint patterns the gels must be "scored." That is to say, the patterns of illuminated bands must be converted into data vectors, whose elements are indicator variables for each of the different bands present across all samples. This scoring is accomplished either manually, if the data are simple, or in most cases by a computer software package, such as *Bionumerics* (Applied Maths, Kortijk, Belgium), that

is specifically designed to score images of community fingerprint gels. If further analysis indicates that a particular band is of interest, the band can be excised from the gel and sequenced for identification in the RDP.

3.1 *Limitations of molecular methods used for microbial ecology*

As with all molecular techniques there are limitations to the methods described that may affect the reliability of community fingerprint data to varying degrees. First, the process of extracting the nucleic acids from the environmental sample may be inefficient and/or biased. Typically, it is difficult to assess the efficiency of the nucleic acid extractions because the total amount of nucleic acid present in the sample is usually unknown. However, protocols for effective nucleic acid extraction specific to the environment from which the sample is taken have been detailed in Akkermans et al. (1996). In addition to the problem of inefficiency, there is evidence that small cells (0.3 to 1.2 μm) are more resistant to cell lysis than larger cells. This indicates that there is the potential that the purified nucleic acid solution obtained may not be representative of the actual community present in the environmental sample. If true, this may lead to systematically biased community fingerprint data. Additionally, there is also variability present in the PCR process. In general, the PCR amplification does not maintain the proportions of the various microbial populations present in the community DNA extracted from the environmental sample. However, this does not usually have a significant effect on community fingerprint data. Another source of variability in the PCR process that is more likely to significantly affect community fingerprint data are chimeric sequences, which result from two strands of DNA from different organisms annealing to one another during one of the amplification cycles. If this happens early in the PCR amplification it can result in a large number of copies of these chimera, and in bands in the community fingerprint, that correspond to no particular organism.

3.2 *Current quantitative methodology for microbial community DNA fingerprint analysis*

Protocols for extraction and purification of microbial community DNA were first developed for aquatic systems. Microbial communities from this environment tend to have fairly simple structure and the community fingerprints only contain a few bands. Visual comparison of community profiles was sufficient in most cases to observe which communities were most similar and which were different. As techniques for the extraction of microbial community DNA from more complex environments with more diverse microbial communities were developed, visual comparison became insufficient for comparison. Researchers then began to apply the techniques of cluster analysis in order to observe which of their sampled communities were most similar. Techniques used include principal components analysis (PCA) (Ranjard et al., 1999), multi-dimensional scaling (MDS) (van Hannen et al., 1999), and hierarchical clustering methods based on similarity indices for binary vectors (Sneath and Sokal, 1973). Unfortunately, while giving some guidance as to which communities were most similar, none of these methods allow researchers to establish conclusive statistical evidence with regard to the specific research questions addressed in their studies. Therefore, we concentrate on

developing proper statistical methodology for identifying microbial populations that vary significantly according to some treatment effect.

4. Notation

We use the following notation for our $n \times d$ binary data matrix $\boldsymbol{X}$.

$$\begin{aligned}
X_{ij}^k &= \begin{cases} 1 & \text{if the } k\text{th microbial population is present in the} \\ & j\text{th sample from the } i\text{th treatment group.} \\ 0 & \text{otherwise} \end{cases} \\
t &= \text{the number of treatment groups. } i = 1, \ldots, t \\
n_i &= \text{the number of samples in the } i\text{th treatment group. } j = 1, \ldots, n_i \\
n &= \sum_{i=1}^{t} n_i \\
d &= \text{the number of variables (i.e., the dimension) } k = 1, \ldots, d
\end{aligned}$$

Marginally, we model $X_{ij}^k \sim \text{Bernoulli}(p_{ik})$ and we estimate the multivariate dependence structure using the within treatment covariance matrix and the between treatment covariance matrix for the sample.

$$\begin{aligned}
\boldsymbol{S_W} &= \frac{1}{n-1} \sum_{i=1}^{t} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_{i\cdot})(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_{i\cdot})' \\
\boldsymbol{S_B} &= \frac{1}{n-1} \sum_{i=1}^{t} n_i (\bar{\boldsymbol{x}}_{i\cdot} - \bar{\boldsymbol{x}}_{\cdot\cdot})(\bar{\boldsymbol{x}}_{i\cdot} - \bar{\boldsymbol{x}}_{\cdot\cdot})' \\
\boldsymbol{S} &= \boldsymbol{S_W} + \boldsymbol{S_B}
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{x}_{ij} &= \text{the } j\text{th sample vector from the } i\text{th treatment group.} \\
\bar{\boldsymbol{x}}_{i\cdot} &= \frac{1}{n_i} \sum_{j=1}^{n_i} \boldsymbol{x}_{ij} \\
\bar{\boldsymbol{x}}_{\cdot\cdot} &= \frac{1}{n} \sum_{i=1}^{t} n_i \bar{\boldsymbol{x}}_{i\cdot} = \frac{1}{n} \sum_{i=1}^{t} \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}
\end{aligned}$$

5. Variable selection

One of the most common approaches in multivariate classification problems is to construct linear discriminating functions $\boldsymbol{f_h}$, $h = 1, \ldots, q \leq \min(t-1, d)$, which maximize $\lambda_h = \frac{\boldsymbol{f_h' S_B f_h}}{\boldsymbol{f_h' S_W f_h}}$ subject to the constraint $\boldsymbol{F' S_W F = I_q}$, where $\boldsymbol{f_h}$ is the h^{th} column of $\boldsymbol{F}$ and

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q$. That is, $\lambda_1, \lambda_2, \ldots, \lambda_q$ are the eigenvalues of $\boldsymbol{S_W}^{-1}\boldsymbol{S_B}$ and $\boldsymbol{f_1}, \boldsymbol{f_2}, \ldots, \boldsymbol{f_q}$ are the corresponding eigenvectors normalized so that $\boldsymbol{F'S_WF} = \boldsymbol{I_q}$ (Hand, 1997). These $\boldsymbol{f_h}$ can also be used, much like some use principal component loadings in the case where groups are not defined *a priori*, to identify a subset of the variables that explain much of the variation in the original d variables associated with the treatment effect. However, in high dimensions, and especially when $d \approx n$ or $d \geq n$, $\boldsymbol{S_W}^{-1}$ can be a poor estimate of the inverse of the population covariance, leading to very inefficient classification and variable selection (Bai and Saranadasa, 1996). For this reason we propose two alternative methods for variable selection in high dimensions.

5.1 *Variable selection using an estimate of* $\boldsymbol{S_W}^{-1}\boldsymbol{S_B}$

Following a similar approach to that detailed above, we assume that $\boldsymbol{S_W}^{-1}\boldsymbol{S_B}$ can be estimated reasonably by $\boldsymbol{W}$, where

$$W[i,j] = \frac{S_B[i,j]}{\sqrt{S_W[i,i]S_W[j,j]}}$$

We then compute the discriminating functions $\boldsymbol{f_h}$, $h = 1, \ldots, q \leq \min(t-1, d)$ as the eigenvectors of $\boldsymbol{W}$, normalized so that $\boldsymbol{F'S_WF} = \boldsymbol{I_q}$. Specifically, we select the subset of variables

$$M = \{X^k | f_{hk} \notin (C_{h,\alpha/2}, C_{h,1-\alpha/2}) \text{ for at least one } h, h = 1, \ldots, q\}$$

where the C_h are constants such that

$$P(f_{hk} < C_{h,\alpha/2}) = P(f_{hk} > C_{h,1-\alpha/2}) = c$$

for each $h = 1, \ldots, q$ and the experimentwise type I error is α. This multiple testing correction is of the type suggested by Westfall and Young (1993). We estimate the C_h empirically by permuting the samples vectors $\boldsymbol{x}_{ij}$ for all i and j and computing $\boldsymbol{f_h}$ $(h = 1, \ldots, q)$ for each permutation. Permuting in this way, we maintain the same covariance structure (i.e., $\boldsymbol{S}$) across all permutations. Therefore, we avoid the distortions brought on by using $\boldsymbol{S_W}^{-1}$ in high dimensions, while still taking the multivariate dependence structure into account in our variable selection criterion. This approach to variable selection attempts to identify a subset of variables that give a large degree of separation between the t treatment groups, but it is limited in that it will most likely not select two highly correlated variables due to the normalization $\boldsymbol{F'S_WF} = \boldsymbol{I_q}$. This restriction is not usually a concern for the analysis of microbial community data since it is of more interest to identify a reasonably sized subset of the observed microbial populations as candidates for further study (e.g., DNA sequencing, genomic analysis, etc.).

5.2 *Variable selection considering each variable individually*

If it is reasonable to assume that $\boldsymbol{S}$ is nearly diagonal, or if we are simply trying to screen for "interesting" variables without worrying about selecting several variables that contain

similar information, we can employ Pearson's χ^2 test of homogeneity as our variable selection criterion.

$$D_k^2 = \frac{\sum_{i=1}^{t} n_i(\bar{x}_{i\cdot}^k - \bar{x}_{\cdot\cdot}^k)^2}{\bar{x}_{\cdot\cdot}^k(1 - \bar{x}_{\cdot\cdot}^k)}$$

Specifically, we select the set of variables

$$M = \{X^k | D_k^2 > C_{k,\alpha}, k = 1, \ldots, d\}$$

where $C_{k,\alpha}$ is defined as the constant such that under the null hypothesis of homogeneity $P(D_k^2 > C_{k,\alpha}) = c$ for all k and the experimentwise type I error is α. Asymptotically, $D_k^2 \xrightarrow{\mathcal{L}} \chi_{t-1}^2$ as $n_i \to \infty$ (Kendall and Stuart, 1961). However, due to the small sample size and the high dimension of the data, and thus the large number of hypotheses being tested, it is unreasonable to assume that $C_{k,\alpha}$ is the $100(1-\alpha)$th percentile of the χ_{t-1}^2 distribution. Instead, we estimate $C_{k,\alpha}$ empirically by permuting the observations for $\boldsymbol{x}^k$ and computing D_k^2 for each permutation. This method will identify a subset of the original d variables for which there is a statistically significant difference between the observed sample proportions of the t different treatment groups, controlling for an experimentwise type I error of α.

6. Classification

6.1 *Classification using a conditionally independent Bernoulli parameterization*

Upon selecting a subset of variables M, we wish to evaluate the variable selection via a classification rule. The probability that a sample $\boldsymbol{X}_{ij}$ is from treatment group g can be expressed as

$$P(G = g|\boldsymbol{X}_{ij}) = P(G = g|X_{ij}^k, \boldsymbol{X}^k \in M)$$

Similarly, the likelihood of $\boldsymbol{X}_{ij}$, given that it comes from treatment group g is denoted

$$P(\boldsymbol{X}_{ij}|G = g) = P(X_{ij}^k, \boldsymbol{X}^k \in M|G = g)$$

If we then assume that the $\boldsymbol{X}_{ij}^k, j = 1, \ldots, n_i$ and $\boldsymbol{X}^k \in M$ are distributed as independent Bernoulli(p_{ik}) random variables we can construct a natural classification rule

$$P(\boldsymbol{X}_{ij}|G = g) = \prod_{\left\{k|\boldsymbol{X}^k \in M\right\}} p_{gk}^{X_{ij}^k}(1 - p_{gk})^{1-X_{ij}^k} \tag{1}$$

Bayes theorem and the reasonable assumption that the prior probabilities $P(G = g) = \frac{1}{t}$ for $g = 1, \ldots, t$ results in the classification of $\boldsymbol{x}_{ij}$ into group g^* if

$$P(G = g^*|\boldsymbol{x}_{ij}) = \max_g P(G = g|\boldsymbol{x}_{ij})$$

Using this approach we estimate the parameters p_{gk}, $g = 1, \ldots, t$ and $\{k|\ \boldsymbol{X}^k \in M\}$ by the corresponding sample proportions $\bar{x}_{g\cdot}^k$, which are the maximum likelihood estimates of the p_{gk} under the Bernoulli parameterization.

6.2 *Classification using conditionally independent logistic regressions*

More frequently in statistical modeling of multivariate categorical response data, generalized linear models are used. Therefore, in order to compare the results of our classification based on the Bernoulli parameterization to a more standard approach, we now fit a logistic regression model, again assuming independence among the $\boldsymbol{X}^k \in M$ conditional on treatment group.

$$\text{logit}(P(G = g|X_{ij})) = \alpha_{g0} + \sum_{\{k|\boldsymbol{X}^k \in M\}} \alpha_{gk} X_{ij}^k + \varepsilon_{gij} \tag{2}$$

$$g = 1 \ldots t, \quad i = 1 \ldots t, \quad j = 1 \ldots n_i$$

Following the same rationale as described for the Bernoulli model (1) we construct classification rule using Bayes rule and assume a non-informative prior. Here we estimate the parameters $\boldsymbol{\alpha}$ using iteratively weighted least squares (IWLS).

7. Application of methodology to microbial community DNA fingerprint data

7.1 *Data*

The described approach for modeling and variable selection was applied to the data from Nakatsu et al. (2000), where the objective of the study was to investigate the impact of different agronomic treatments on the microbial community structure of corn rhizosphere. Corn plants were grown at the Purdue University Agronomy Research Center in disturbed (plowed) and undisturbed (no-till) soils with a 24 year history of growth as a monoculture crop (corn only) or two crops grown in annual rotation (corn and soybean). Rhizosphere soils were sampled during early developmental stages and a community fingerprint was produced for each sample by DGGE of PCR amplified 16S rDNA from the soil extracted DNA. While there was the potential for very high-dimensional data on the order of $d = 10,000$, we use the $d = 84$ distinct microbial populations that were identified across all $n = 89$ samples ($n_1 = 23, n_2 = n_3 = n_4 = 22$).

7.2 *Variable selection*

We consider both of the proposed variable selection criteria in turn. We first employ the variable selection criteria described in section 5.1 and for simplicity take $q = t - 1 = 3$ and $\alpha = 0.05$. We select only 1 variable, $M_1 = \{X^{13}\}$, based on $C_{h,\alpha/2}$ and $C_{h,1-\alpha/2}$, $h = 1, \ldots, q$ estimated from 10,000 permutations of the data. Table 1 displays the proportion of samples in each treatment for which the selected microbial population (i.e., variable) was present. Clearly, this one binary variable alone will not be sufficient to classify observations into 4 different treatment groups. However, we observe that $\bar{x}_{1\cdot}^{13} = 1.00$, $\bar{x}_{2\cdot}^{13} = \bar{x}_{3\cdot}^{13} = 0.00$ and $\bar{x}_{4\cdot}^{13} = 0.14$, which indicates that a sample which contains the microbial population associated with X^{13} is very likely to have come from treatment 1.

Alternatively, using the variable selection criteria described in section 5.2 and taking $\alpha =$

0.05, a set of 28 variables is selected.

$$\begin{aligned} M_2 = \{&X^9, X^{11}, X^{12}, X^{13}, X^{14}, X^{16}, X^{19}, X^{27}, X^{32}, X^{34}, \\ &X^{36}, X^{39}, X^{40}, X^{41}, X^{42}, X^{43}, X^{45}, X^{46}, X^{48}, \\ &X^{49}, X^{51}, X^{53}, X^{54}, X^{55}, X^{56}, X^{61}, X^{82}, X^{84}\} \end{aligned}$$

Table 1 displays the proportion of samples in each treatment for which the selected microbial populations (i.e., variables) were present.

We observe that most of the 28 selected variables have $\bar{x}_{i\cdot}^k = 0.00$ for at least one $i = 1, \ldots, 4$. This result is somewhat expected since this criterion selects variables for which there is a statistically significant difference between the observed sample proportions of the 4 different treatment groups. It is natural then to select variables corresponding to microbial populations that are present in a large proportion of the samples from at least one treatment and absent in samples from the remaining treatments.

7.3 *Cross-validation*

The two classification rules described in section 6 were employed to validate the subsets of variables selected, M_1 and M_2. Table 2 details the results of a cross-validation. That is to say, we re-estimate the parameters of the models for each observation we are aiming to classify leaving that observation out of the calculations.

As expected we observe that using M_1, both classification rules correctly classify all samples from treatment 1 (i.e., the monoculture/plow treatment), but none of the others. We also observe that using M_2, both classification rules correctly classify more than 85% of the samples, with our Bernoulli classification rule outperforming the more standard logistic classification rule.

8. Summary

In our analysis of the microbial community data we have illustrated the effectiveness of our variable selection methodology for relatively high-dimensional multivariate binary data. However, our analysis also raises a number of issues. The variable selection method from section 5.1 appears to be very restrictive for a number of reasons. First, due to the large number of tests (i.e., dq) and the multiple testing correction, the criterion requires that values of f_{hk} for any selected variable be in the extreme tails of permutation distribution. Secondly, as addressed in section 5.1, due to the normalization $\boldsymbol{F'S_WF} = \boldsymbol{I_q}$, this method is not likely to select two highly correlated variables. And finally, by design, the variation explained by $\boldsymbol{f_1}$ is greater than that explained by $\boldsymbol{f_2}$ and so on. Therefore, one can think of many ways to adjust the multiple testing correction to make a more reasonable variable selection, but no such adjustment would have affected the variable selection made for the Nakatsu et al. (2000) data.

In truly high-dimensional problems (i.e., $d \approx 10,000$) the property of restrictive variable

selection might be advantageous in application. For example, in our application we select 1 out of 84 variables in the Nakatsu et al. (2000) data using the method of section 5.1 and 28 out of the 84 variables using the method of section 5.2. In a situation where $d = 10,000$, selecting the same proportion of variables, as in our example, would yield $\frac{10000}{84} \approx 119$ variables selected by the method of section 5.1 and $\frac{280000}{84} \approx 3333$ variables selected by the method of section 5.2. This extrapolation indicates that for truly high-dimensional data the method of section 5.1 might be more appropriate. However, the method of section 5.2 is not unreasonable in high dimensions because as the dimension, and consequently the number of hypotheses tested, increase the comparisonwise error rate becomes extremely small, resulting in a much more restrictive variable selection criterion.

In addition, we realize that the assumption of conditional independence made, in order to construct our simple classification rules, may have a significant effect on our ability to evaluate subsets of dependent variables via classification. Nevertheless, this assumption is supported by current ecological theory for the microbial communities in rhizosphere soil (Coyne, 1999). In situations where one does have dependent subsets of selected variables, one might modify the classification rules in order to account for the dependence structure among the selected variables. Due to the potential for high-dimensional subsets of selected variables, this might best be accomplished using latent variable models to reduce the number of estimated parameters.

Clearly, the framework of our variable selection methodology, is not limited to microbial community characterization. The number of sources of high-dimensional data continue to increase, especially in the biological sciences where advances in molecular technology and the ever increasing interest in functional genomics has led to the production of massive data, some of it being binary. Therefore, our methodology could be applied to the problem of variable selection for high-dimensional binary data in many such fields, as well for continuous data with some modification.

9. Acknowledgments

This work was supported by a grant from the USDA National Research Initiative Soils and Soil Biology Program (98-35107-6389) to Sylvie M. Brouder, Cindy H. Nakatsu, and R.W. Doerge. Jayson D. Wilbur is supported from a Purdue Research Foundation grant to R. W. Doerge. We would also like to thank Judy Lindell and the entire tillage group for technical assistance and continual support.

References

Akkermans, A. D. L., van Elsas, J. D. and de Bruijn, F. J. (1996). *Molecular microbial ecology manual.* Kluwer Academic Publishers, Norwell, MA.

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension by an example of a two sample problem. *Statistica Sinica* **6**, 311–329.

Chiarini, L., Bevivino, A., Dalmastri, C., Nacamulli, C. and Tabacchioni, S. (1998). Influence of plant development, cultivar and soil type on microbial colonization of maize roots. *Applied Soil Ecology* **8**, 11–18.

Coyne, M. (1999). *Soil Microbiology: an exploratory approach.* Delmar Publishers.

Hand, D. J. (1997). *Construction and Assessment of Classification Rules.* John Wiley & Sons Ltd.

Kendall, M. and Stuart, A. (1961). *Kendall's Advanced Theory of Statistics Volume 2: Classical Inference and Relationship (Fifth Edition).* Hafner Publishing Company.

Maidak, B. L., Cole, J. R., Lilburn, T. G., Parker, C. T., J., Saxman, P. R., Farris, R. J., Garrity, G. M., Olsen, G. J., Schmidt, T. M. and Tiedje, J. M. (2001). The rdp-ii (ribosomal database project). *Nucleic Acid Research* **29**, 173–174.

Muyzer, G. and Smalla, K. (1998). Application of denaturing gradiaent gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE) in microbial ecology. *Antonie van Leeuwenhoek - International Journal of General and Molecular Microbiology* **73**, 127–141.

Nakatsu, C. H., Brouder, S. M., Wilbur, J. D., Wanjau, F. and Doerge, R. W. (2000). Impact of tillage and crop rotation on corn development and its associated microbial community. In *Proceedings of 15th Conference of the International Soil Tillage Research Organization (ISTRO).*

Pace, N. R., Stahl, D. A., Lane, D. J. and Olsen, G. J. (1986). The analysis of natural microbial populations by ribosomal rna sequences. *Adv Microb Ecol* **9**, 1–55.

Ranjard, L., Nazaret, S., Gourbière, F., Thioulouse, J., Philippe, L. and Richaume, A. (1999). A soil micro scale study to reveal the heterogenetiy of Hg(II) on indigenous bacteria by quantification of adapted phenotypes and anlysis of community DNA fingerprints. *FEMS Microbiology Ecology* **31**, 107–115.

Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A. and Arnheim, N. (1985). Enzymatic amplification of Beta-globin genomic sequences and restriction site analysis for diagnosis of sickle-cell anemia. *Science* **230**, 1350–1354.

Sneath, P. H. A. and Sokal, R. R. (1973). *Numerical Taxonomy.* W. H. Freeman.

van Hannen, E. J., Zwart, G., van Agterveld, M. P., Gons, H. J., Ebert, J. and Laanbroek, H. J. (1999). Changes in bacterial and eukaryotic community structure after mass lysis of *Filamentous Cyanobacteria* associated with viruses. *Applied and Environmental Microbiology* **65**, 795–801.

West, T. D., Griffith, D. R., Steinhardt, G. C., Kladivko, E. J. and Parsons, S. D. (1996). Effect of tillage and rotation on agronomic performance of corn and soybean: Twenty-year study on dark silty clay loam soil. *Journal of Production Agriculture* **9**, 241–248.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing.* John Wiley & Sons Inc.

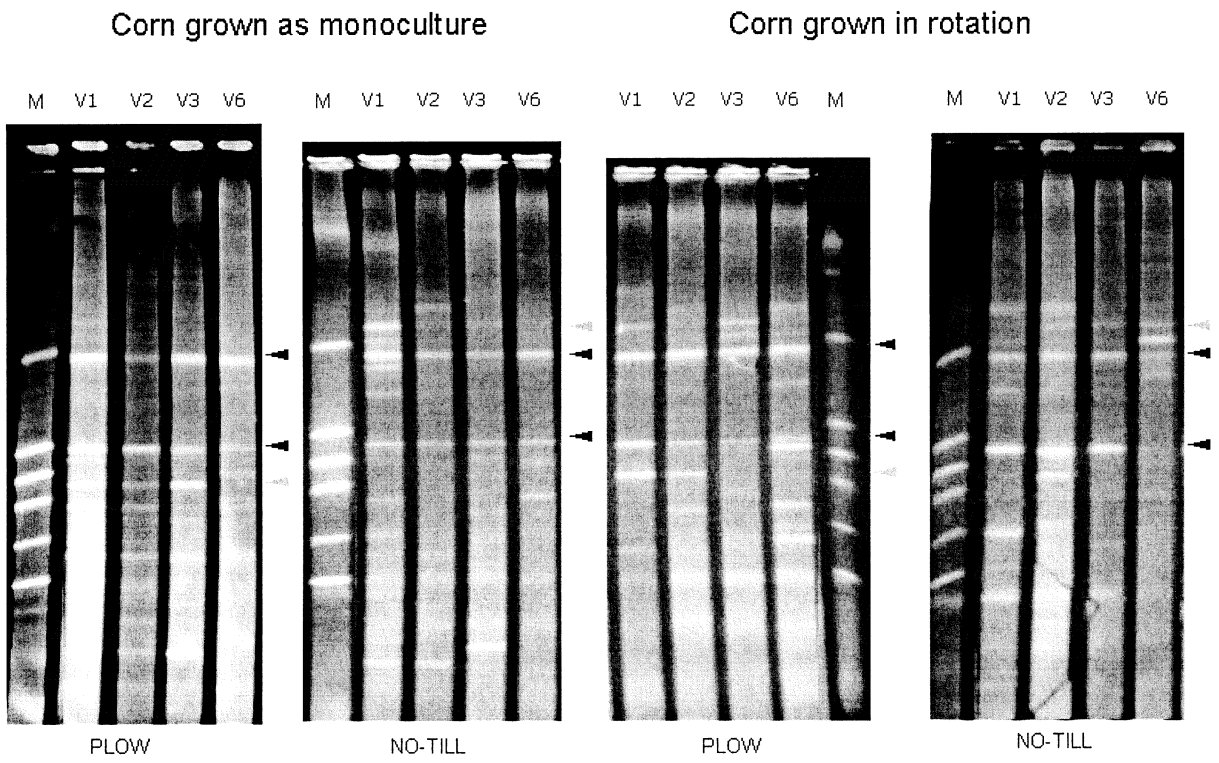

Figure 1. Representative characteristic profiles from 4 agronomic treatments are shown above. Each vertical column or "lane" represents the profile for one microbial community. Within each profile the pattern of illuminated bands reveals distinct fingerprint patterns which can be used to distinguish microbial community structure (i.e., which populations of microorganisms are present in the community). The tillage practice for each sample (i.e., plow or no-till) is listed above each block of lanes, or "gel." The rotation practice for each sample (i.e., monoculture or rotation) is listed below the gels. The growth stage of the associated plant for each sample listed above each lane (i.e., V1, V2, V3 and V6) and lane M on each gel is a "marker lane" common to all gels to enable between gel comparisons. The black arrows denote some bands common to all agronomic treatments and the grey arrows denote some bands present only in samples from specific treatments.

k	$\bar{x}^k_{1.}$	$\bar{x}^k_{2.}$	$\bar{x}^k_{3.}$	$\bar{x}^k_{4.}$	k	$\bar{x}^k_{1.}$	$\bar{x}^k_{2.}$	$\bar{x}^k_{3.}$	$\bar{x}^k_{4.}$
9	0.2609	1.0000	0.3182	0.5455	42	0.0000	0.0000	0.2273	0.0455
11	0.0000	0.0909	0.0000	0.0000	43	0.0000	0.1364	0.0000	0.3182
12	0.0000	0.4545	0.0000	0.3636	45	0.0000	0.0000	0.2727	0.0000
• 13	1.0000	0.0000	0.0000	0.1364	46	0.0870	0.4545	0.1364	0.0000
14	1.0000	0.6818	1.0000	0.7727	48	0.0435	0.5000	0.0000	0.0000
16	0.0000	0.0000	0.0000	0.2273	49	0.0000	0.0000	0.0000	0.2727
19	0.1739	0.4091	0.0000	0.0000	51	0.0000	0.0000	0.0000	0.2273
27	0.0000	0.0909	0.0000	0.0000	53	0.0000	0.2727	0.0000	0.0455
32	0.7391	0.3182	0.0000	0.1364	54	0.8696	0.0000	0.6364	0.0000
34	0.0000	0.0000	0.7727	0.4091	55	0.0000	0.0000	0.0000	0.2727
36	0.4783	0.1818	0.0000	0.0000	56	0.0000	0.1364	0.0000	0.0000
39	0.0870	0.3636	0.0000	0.0000	61	0.0000	0.0909	0.0000	0.0000
40	0.0870	0.1818	0.0000	0.4091	82	0.0000	0.0909	0.0000	0.0000
41	0.0000	0.0909	0.0000	0.0000	84	0.4783	0.0000	0.6364	0.3182

Table 1

Applying the variable selection methodology of section 5.2 to the Nakatsu data we select the 28 variables displayed above along with the proportion of samples in each treatment for which the selected microbial populations (i.e., variables) were present. The single variable selected using the methodology of section 5.1 is X^{13}, highlighted by a dot to the left of the variable number in the table.

		Treatment				
		Monoculture		Rotated		
		Plow	No-Till	Plow	No-Till	
Subset	Rule	1	2	3	4	Total
M_1	Bernoulli	23	0	0	0	23
M_1	Logistic	23	0	0	0	23
M_2	Bernoulli	22	19	22	19	82
M_2	Logistic	22	20	20	15	77
	Observations	23	22	22	22	89

Table 2

Number of correctly classified samples in the cross-validations described in section 7.3. M_1 is the subset of variables selected using the method described in section 5.1 and M_2 is the subset of variables selected using the method described in section 5.2