Variable Selection in High-Dimensional Multivariate Binary Data with Application to the Analysis of Microbial Community DNA Fingerprints

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

[1]Department of Statistics, Purdue University, West Lafayette, Indiana 47907-1399, U.S.A.
[2]Computational Genomics, Purdue University, West Lafayette, Indiana 47907, U.S.A.
[3]Indian Statistical Institute, Calcutta, India
[4]Department of Agronomy, Purdue University, West Lafayette, Indiana 47907-1150, U.S.A.
*email: doerge@purdue.edu

SUMMARY. In order to understand the relevance of microbial communities on crop productivity, the identification and characterization of the rhizosphere soil microbial community is necessary. Characteristic profiles of the microbial communities are obtained by denaturing gradient gel electrophoresis (DGGE) of polymerase chain reaction (PCR) amplified 16S rDNA from soil extracted DNA. These characteristic profiles, commonly called community DNA fingerprints, can be represented in the form of high-dimensional binary vectors. We address the problem of modeling and variable selection in high-dimensional multivariate binary data and present an application of our methodology in the context of a controlled agricultural experiment.

KEY WORDS: Classification; DNA fingerprints; High-dimensional data; Microbial communities; Multivariate binary data; Permutation tests; Variable selection.

1. Introduction

Studies have found that, when a single crop species, such as corn, is grown continually without the rotation of other crops, yield decline occurs (Dick and Van Doren 1985; Cronkston, Kurle, and Lueschen, 1988; Griffith et al., 1988). Furthermore, analysis of yields over the long term suggests that there is 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 components of the ecosystem to biotic phenomena mediated by the microbial community present in the rhizosphere soil (Bevivino et al., 1998; Chiarini et al., 1998; Turco et al., 1990). The rhizosphere is the portion of the soil volume in intimate contact with the growing root system.

Historically, there have been limitations in the molecular methodology available to examine the general ecology of microbial systems, particularly those in soil. Recently though, a number of molecular methods have been developed that enable the direct analysis of microbial populations in soil (Akkermans, van Elsas, and de Bruijn, 1996; Torsvik et al., 1998). For example, characteristic profiles of the microbial communities, commonly called community DNA fingerprints, can be produced by denaturing gradient gel electrophoresis (DGGE) of 16S rDNA obtained by polymerase chain reaction (PCR) amplification of the DNA extracted from the rhizosphere soil. Figure 1 illustrates an example of representative profiles from four agronomic treatments.

These microbial community fingerprints can be represented in the form of binary vectors, which have the potential to be of very high dimension because the number of bacterial types in some environmental soil samples have been estimated to be on the order of 10,000 (Torsvik, Sorheim, and Goksoyr, 1996). The quantitative methodology currently used for the analysis of these microbial community fingerprints falls exclusively within the scope of exploratory data analysis. While some researchers use principal components (Ranjard et al., 1999; Miethling et al., 2000) or multidimensional scaling (van Hannen et al., 1999a,b; Iwamoto et al., 2000) to find groups of similar observations, the vast majority use hierarchical clustering algorithms based on similarity indices for binary vectors (Jaccard, 1908; Dice, 1945). Some researchers even avoid quantitative analysis altogether by drawing conclusions based on visual comparison of the microbial community fingerprints (ben Omar and Ampe, 2000; Felske et al., 2000; Tannock et al., 2000). Unfortunately, all these exploratory techniques do little to establish conclusive statistical evidence with regard to the specific research questions addressed in their studies. Toward this end, we concentrate on the characterization of the microbial community structure in terms of statistical models and aim to identify a subset of the variables (i.e., microbial populations) that contribute to the variation in the system associated with the treatment effects.

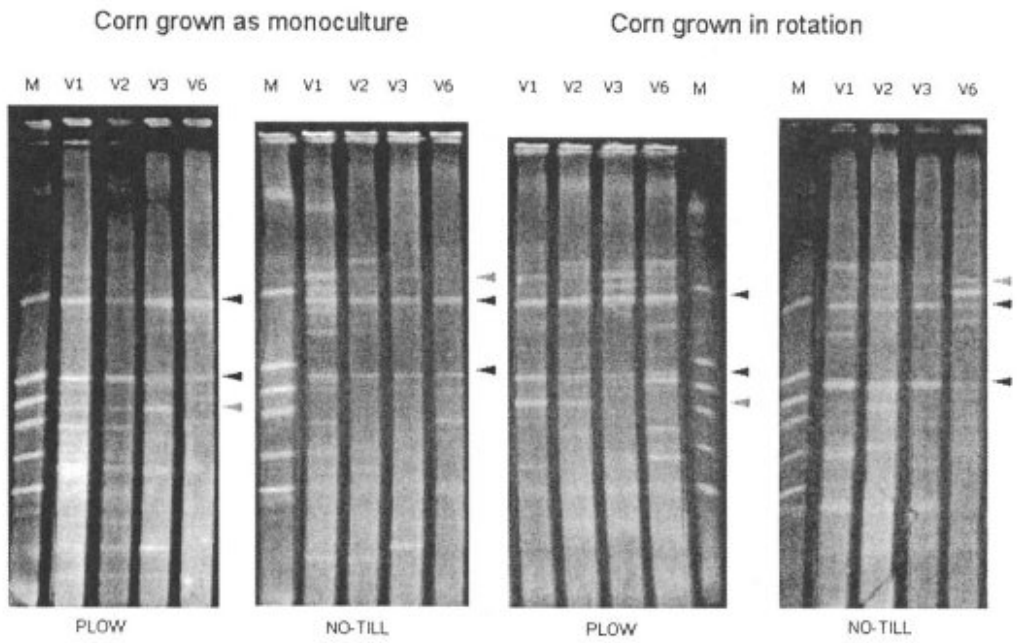

Figure 1. Representative characteristic profiles from four agronomic treatments. Each vertical column, or lane, represents the profile for one microbial community. Within each profile, the pattern of illuminated bands reveals distinct fingerprint patterns, which can be used to distinguish microbial community 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 below each block of lanes, or gel. The rotation practice for each sample (i.e., monoculture or rotation) is listed above 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, which enables 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.

While there is an increasing number of algorithmic approaches for variable selection in high-dimensional data that fall under the umbrella of data mining, little work based in probabilistic modeling has been done, and none addresses the specific issues associated with binary data. And while there is limited work (Nuamah, Qu, and Amini, 1996; Sohn, 1999) that addresses variable selection in relatively low-dimensional correlated binary regression, none addresses the influence of high dimension in their methodology. For this reason, we propose new variable selection criteria specific to high-dimensional multivariate binary data and apply them to microbial community fingerprint data.

2. Notation

We use the following notation for the $n \times d$ binary data matrix $\boldsymbol{X}$. Let X_{ij}^k be an indicator for the presence of the kth microbial population in the jth sample from the ith treatment group. Let t denote the number of treatment groups, d the number of variables (i.e., the dimension), n the number of samples, and n_i the number of samples from treatment i. 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 as

$$\boldsymbol{S_W} = \frac{1}{n-1} \sum_{i=1}^{t} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_{i\cdot})(\boldsymbol{x}_{ij} - \boldsymbol{x}_{i\cdot})'$$

$$\boldsymbol{S_B} = \frac{1}{n-1} \sum_{i=1}^{t} n_i (\bar{\boldsymbol{x}}_{i\cdot} - \bar{\boldsymbol{x}}_{\cdot\cdot})(\bar{\boldsymbol{x}}_{i\cdot} - \bar{\boldsymbol{x}}_{\cdot\cdot})',$$

where $\boldsymbol{x}_{ij}$ denotes the jth sample vector from the ith treatment group, $\bar{\boldsymbol{x}}_{i\cdot}$ is the ith treatment mean vector, and $\bar{\boldsymbol{x}}_{\cdot\cdot}$ is the grand mean vector.

3. 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 $\lambda_h = (\boldsymbol{f_h'} \boldsymbol{S_B} \boldsymbol{f_h})/(\boldsymbol{f_h'} \boldsymbol{S_W} \boldsymbol{f_h})$ subject to the constraint $\boldsymbol{F'} \boldsymbol{S_W} \boldsymbol{F} = \boldsymbol{I_q}$, where $\boldsymbol{f_h}$ is the hth column of $\boldsymbol{F}$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q$, i.e., $\lambda_1, \lambda_2, \ldots, \lambda_q$ 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 $\boldsymbol{F'} \boldsymbol{S_W} \boldsymbol{F} = \boldsymbol{I_q}$ (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.

3.1 *Variable Selection Using an Estimate of* $\boldsymbol{S_W^{-1}S_B}$

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

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

The effects of this assumption on variable selection are explored via a simulation study in Section 7. 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 $\boldsymbol{F'S_WF = I_q}$. Specifically, we select the subset of variables

$$M = \left\{ X^k \mid f_{hk} \notin \left(C^k_{h,\alpha/2}, C^k_{h,1-\alpha/2} \right) \text{ for at least one } h, \right.$$
$$\left. h = 1, \ldots, q \right\},$$

where the C^k_h are constants such that

$$P\left(f_{hk} < C^k_{h,\alpha/2}\right) = P\left(f_{hk} > C^k_{h,1-\alpha/2}\right) = c$$

for $k = 1, \ldots, d$ and $h = 1, \ldots, q$ and the experimentwise type I error is α. We estimate the C^k_h empirically by permuting the sample vectors $\boldsymbol{x}_{ij}$ for all i and j and computing $\boldsymbol{f_h}$ $(h = 1, \ldots, q)$ for each permutation. This multiple testing correction is of the general type suggested by Westfall and Young (1993). Permuting in this way, we maintain the same covariance structure (i.e., $\boldsymbol{S = S_W + S_B}$) across all permutations. Therefore, because we permute in this way and because we still use the full $\boldsymbol{S_W}$ matrix in the normalization $\boldsymbol{F'S_WF = I_q}$, we are able to 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.

Approaching variable selection in this manner attempts to identify a subset of variables that give a large degree of separation between the t treatment groups while making adjustments for the within-treatment dependence structure. Such adjustments for the within-treatment dependence structure are not always preferable in the analysis of microbial community data because the researcher is usually more interested in identifying a reasonably sized subset of the observed microbial populations as candidates for further study (e.g., DNA sequencing, genomic analysis, etc.).

3.2 *Variable Selection Considering Each Variable Individually*

If we are simply trying to screen for interesting variables without worrying about the dependence between the variables, we can use the following test statistic as our variable selection criterion:

$$D^2(k) = \frac{S_B[k,k]}{S_W[k,k]} = \frac{\sum_{i=1}^{t} n_i \left(\bar{x}^k_{i\cdot} - \bar{x}^k_{\cdot\cdot}\right)^2}{\sum_{i=1}^{t} n_i \bar{x}^k_{i\cdot}\left(1 - \bar{x}^k_{i\cdot}\right)}.$$

Specifically, we select the set of variables

$$M = \left\{ X^k : D^2(k) > C^k_\alpha, k = 1, \ldots, d \right\},$$

where C^k_α is defined as the constant such that, under the null hypothesis of homogeneity, $P(D^2(k) > C^k_\alpha) = c$ for all k and the experimentwise type I error is α. We estimate C^k_α empirically by permuting the sample vectors $\boldsymbol{x}_{ij}$ as before, maintaining the same covariance structure (i.e., $\boldsymbol{S}$) across all permutations, and computing the vector D^2 for each permutation. This method will identify the 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.

4. Classification

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

$$P(i = g \mid \boldsymbol{X}_{ij} = \boldsymbol{x}_{ij}) = \frac{P(\boldsymbol{X}_{ij} = \boldsymbol{x}_{ij} \mid i = g)P(i = g)}{\sum_g P(\boldsymbol{X}_{ij} = \boldsymbol{x}_{ij} \mid i = g)P(i = g)}.$$

So we will classify sample $\boldsymbol{X}_{ij}$ into group g^* if

$$P(i = g^* \mid \boldsymbol{X}_{ij} = \boldsymbol{x}_{ij}) = \max_g P(i = g \mid \boldsymbol{X}_{ij} = \boldsymbol{x}_{ij}).$$

4.1 *Classification Using a Conditionally Independent Bernoulli Parameterization*

If we assume that the X^k_{ij} are distributed as independent Bernoulli(p_{ik}) random variables for $j = 1, \ldots, n_i$ and $\boldsymbol{X}^k \in M$, then

$$P(\boldsymbol{X}_{ij} = \boldsymbol{x}_{ij} \mid i = g) = \prod_{k \in K} p_{gk}^{x^k_{ij}} (1 - p_{gk})^{1 - x^k_{ij}}, \quad (1)$$

where $K = \{k : \boldsymbol{X}^k \in M\}$. To estimate $P(i = g \mid \boldsymbol{X}_{ij} = \boldsymbol{x}_{ij})$, we make the reasonable assumption that $P(i = g) = 1/t$ for $g = 1, \ldots, t$ and estimate the parameters p_{gk}, $g = 1, \ldots, t$ and $k \in K$ by the corresponding sample proportions $\bar{x}^k_{g\cdot}$, which are the maximum likelihood estimates of the p_{gk} under the Bernoulli parameterization.

4.2 *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 with more a standard approach, we now fit a logistic regression model, again assuming independence among the $\boldsymbol{X}^k \in M$, conditional on treatment group,

$$\text{logit}(P(i = g \mid \boldsymbol{X}_{ij} = \boldsymbol{x}_{ij})) = \alpha_{g0} + \sum_{k \in K} \alpha_{gk} x^k_{ij} + \varepsilon_{gij}, \quad (2)$$

where $g = 1, \ldots, t$, $i = 1, \ldots, t$, and $j = 1, \ldots, n_i$. Here we estimate the parameters α using iteratively weighted least squares (IWLS) (McCullagh and Nelder, 1989; Venables and Ripley, 1999).

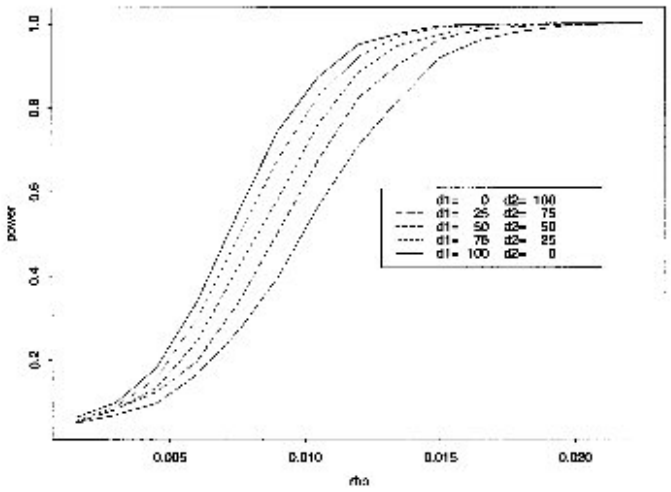

Figure 2. Power curves for the test of diagonality of $\boldsymbol{S_W}$ with respect to the alternative hypothesis of an intraclass correlation structure with parameter ρ. The intraclass correlation ρ varies across the horizontal axis and the different curves correspond to different $d = 100$-dimensional Bernoulli probability vectors. $d1$ corresponds to the number of Bernoulli(0.05) variables and $d2 = 100 - d1$ corresponds to the number of Bernoulli(0.50) variables in each of the $n = 100$ samples for each simulated data. For each combination of $d1, d2$, and ρ, we simulated 5000 data with $t = 4$ treatments and $n_1 = n_2 = n_3 = n_4 = 25$, and calculated the power of the test.

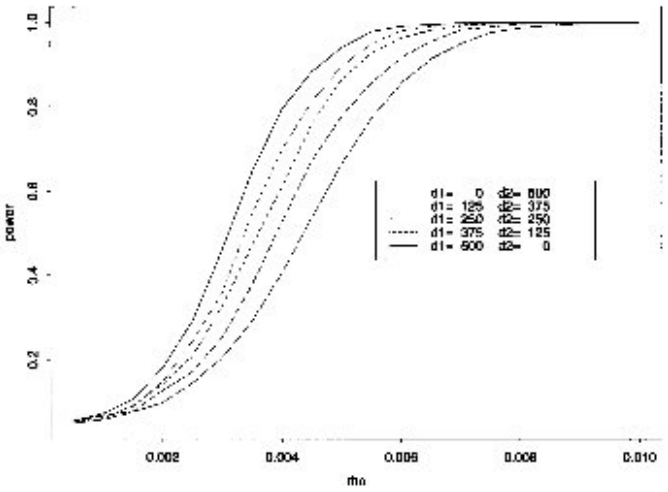

Figure 3. Power curves for the test of diagonality of $\boldsymbol{S_W}$ with respect to the alternative hypothesis of an intraclass correlation structure with parameter ρ. The intraclass correlation ρ varies across the horizontal axis and the different curves correspond to different $d = 500$-dimensional Bernoulli probability vectors. $d1$ corresponds to the number of Bernoulli(0.05) variables and $d2 = 500 - d1$ corresponds to the number of Bernoulli(0.50) variables in each of the $n = 100$ samples for each simulated data. For each combination of $d1, d2, \rho$ we simulated 5000 data with $t = 4$ treatments and $n_1 = n_2 = n_3 = n_4 = 25$, and calculated the power of the test.

5. Testing the Assumption of Conditional Independence

Because both of the models we use to construct classification rules assume independence among the selected variables conditional on the treatment group, it is important to test this assumption. Typically, one method of testing for independence among the selected variables conditional on treatment group is to assume that the binary variables are dichotomies of latent multivariate normal random variables and then use a likelihood ratio test for independence under the multivariate normal assumption (i.e, the determinant of the sample correlation matrix). However, due to the potential for relatively high-dimensional M, leading to singularity in the sample correlation matrix with high probability, we propose the following test as an alternative.

5.1 *Testing Diagonality of* $\boldsymbol{S_W}$

We take the ratio of the sum of the absolute value of the off-diagonal elements to the diagonal elements of $\boldsymbol{S_W}$ for $\boldsymbol{X}^k \in M$ to develop a test statistic, $T_W(\boldsymbol{x})$, for testing the diagonality of $\boldsymbol{S_W}$ as

$$T_W(\boldsymbol{x}) = \frac{\sum_{k_1 \in K} \sum_{\{k_2 \; k_2 < k_1, k_2 \in K\}} |\boldsymbol{S_W}[k_1, k_2]|}{\mathrm{tr}(\boldsymbol{S_W})}.$$

In order to estimate the distribution of T_W under the null hypothesis of diagonality, the observations for each $\boldsymbol{X^k} \in M$ are permuted independently within each treatment group. The within-treatment covariance matrix and test statistic T_W are then computed for each of the permutations. We reject the hypothesis of diagonality if $T_W(\boldsymbol{x})$ is greater than the $100(1-\alpha)$th percentile of the distribution of T_W under the null hypothesis.

5.2 *Power Calculations*

The power of the proposed test of diagonality is investigated using the algorithm of Ennrich and Piedmonte (1991) to simulate data from the alternative hypotheses, $\boldsymbol{S_W} = (1-\rho)\boldsymbol{I} + \rho\boldsymbol{J}$, where $\boldsymbol{J}$ is the matrix of ones. For these simulations, we take $t = 4$ and $n_1 = n_2 = n_3 = n_4 = 25$. We estimate the power of the test for $\rho > 0$ using five different d-dimensional Bernoulli probability vectors held constant across the t treatments. Our findings for $d = 100$ and $d = 500$ are presented in Figures 2 and 3, respectively.

We observe in the case of $d = 100$ that the power of the test is greater than 0.90 for $\rho > 0.015$ regardless of the composition of the vector of Bernoulli probabilities. In the case of $d = 500$, the power of the test is greater than 0.80 for $\rho > 0.0055$ and greater than 0.90 for $\rho > 0.0065$ regardless of the composition of the vector of Bernoulli probabilities. Therefore, we regard this test of the hypothesis of diagonality of $\boldsymbol{S_W}$ as very powerful against the alternative hypothesis of an intraclass correlation structure.

In addition, comparison of the results for $d = 100$ and $d = 500$ indicates that the shapes of the power curves are very similar regardless of dimension. We also observe that the shapes of the power curves are similar regardless of the composition of the Bernoulli probability vector, and as the proportion of Bernoulli(0.05) variables increases, the steepness of the curve also increases. However, we observe

Table 1

Application of the variable selection methodology of Section 3.2 to the Nakatsu et al. (2000) data. Nineteen variables are selected and displayed along with the proportion of samples in each treatment for which the selected microbial populations (i.e., variables) were present. The three variables selected using the methodology of Section 3.1 are denoted by a dot.

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

that, for a fixed value of ρ, the power of the test is much greater for $d = 500$ than for $d = 100$. We anticipate that these trends will persist in higher dimensions.

6. Application of Methodology to Microbial Community DNA Fingerprint Data

6.1 *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 25-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 $d = 10{,}000$, here only $d = 84$ distinct microbial populations were identified across all $n = 89$ samples. The distribution of the samples across the four treatment groups is $n_1 = 23, n_2 = n_3 = n_4 = 22$, where the treatments are (1) corn grown in monoculture in plowed soil, (2) corn grown in monoculture in undisturbed (no-till) soil, (3) corn grown in rotation with soybean in plowed soil, and (4) corn grown in rotation with soybean in undisturbed (no-till) soil.

Table 2

Number of correctly classified samples in the cross-validations described in Section 6.3. The treatments are (1) corn grown in monoculture in plowed soil, (2) corn grown in monoculture in no-till soil, (3) corn grown in rotation with soybean in plowed soil, (4) corn grown in rotation with soybean in no-till soil.

		Treatment				
Subset	Rule	1	2	3	4	Total
M_1	Bernoulli	23	22	14	9	68
M_1	Logistic	20	22	14	9	65
M_2	Bernoulli	22	18	22	17	79
M_2	Logistic	22	18	18	14	72
	Observations	23	22	22	22	89

6.2 *Variable Selection*

Each of the proposed variable selection criteria is considered in turn. We first employ the variable selection criteria described in Section 3.1 and, for simplicity, take $q = t - 1 = 3$ and $\alpha = 0.05$. We select a subset of three variables, $M_1 = \{X^{13}, X^{34}, X^{54}\}$, based on $C^k_{h,\alpha/2}$ and $C^k_{h,1-\alpha/2}$, $h = 1, \ldots, q$, estimated from 10,000 permutations of the data. Alternatively, using the variable selection criteria described in Section 3.2 and taking $\alpha = 0.05$, a subset of 19 variables is selected as

$$M_2 = \{X^9, X^{12}, X^{13}, X^{14}, X^{19}, X^{32}, X^{34}, X^{36}, X^{39}, X^{40}, X^{43}, X^{45}, X^{46}, X^{48}, X^{49}, X^{53}, X^{54}, X^{55}, X^{84}\}.$$

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 19 selected variables have $\bar{x}^k_{i\cdot} = 0.00$ for at least one $i = 1, \ldots, 4$. This result is somewhat expected because both criteria select variables for which there is a significant separation between the four 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.

6.3 *Classifications*

The two classification rules described in Section 4 were employed to validate the subsets of variables, M_1 and M_2, selected in Section 6.2. Table 2 details the results of a cross-validation of the observations in the sample using only the selected variables. 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.

Using M_1, both classification rules correctly classify more than 70% of the samples, and using M_2, both classification rules correctly classify more than 80% of the samples, with our Bernoulli classification rule doing slightly better than the more standard logistic classification rule. The Bernoulli classification rule may perform slightly better than the logistic classification rule in this case because all the variables involved are in the binary scale. Typically, logistic models are used to model binary responses (or proportions) as a function of predictor variables or covariates measured on a continuous scale.

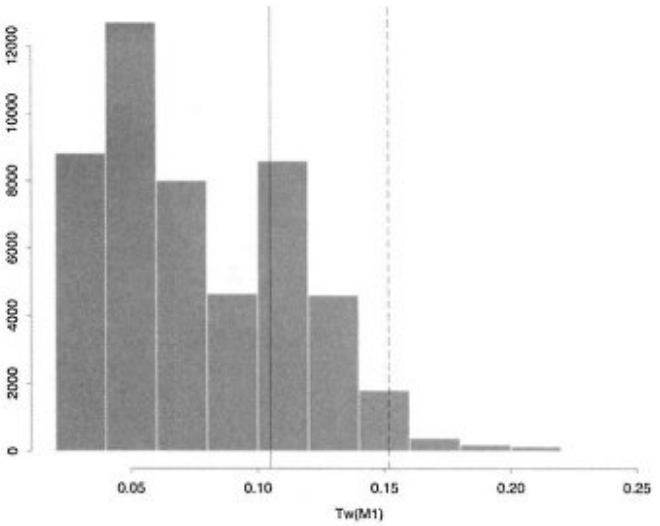

Figure 4. A histogram of the distribution of test statistic T_W under the null hypothesis of independence between the selected bands within treatment/growth-stage groups. The test statistic is the ratio of the absolute value of the off-diagonal elements of the within-treatment sample covariance matrix for the three variables in M_1. The dotted line denotes the 95th percentile of the distribution, 0.1517, and the solid line denotes $T_W(\boldsymbol{x}) = 0.1050$, the test statistic for the data. Therefore, we do not reject the hypothesis of independence between the selected bands within treatment groups at the $\alpha = 0.05$ level.

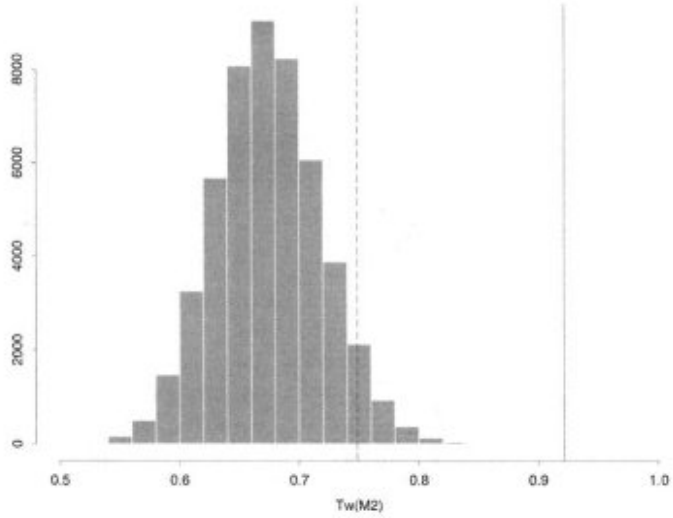

Figure 5. A histogram of the distribution of test statistic T_W under the null hypothesis of independence between the selected bands within treatment/growth-stage groups. The test statistic is the ratio of the absolute value of the off-diagonal elements of the within-treatment sample covariance matrix for the 19 variables in M_2. The dotted line denotes the 95th percentile of the distribution, 0.7486, and the solid line denotes $T_W(\boldsymbol{x}) = 0.9216$, the test statistic for the data. Therefore, we reject the hypothesis of independence between the selected bands within treatment groups at the $\alpha = 0.05$ level.

6.4 *Tests for Diagonality of S_W for $X^k \in M$*

Using the method developed in Section 5.1, we test for diagonality among the selected variables conditional on treatment group with $\alpha = 0.05$. The null hypothesis of diagonality for M_1 (Figure 4) is not rejected, but we do reject the null hypothesis for M_2 (Figure 5). These results are somewhat expected due the way the variables are selected by each method. The variables in M_1 are selected using a method designed to select a subset of variables that provides a large degree of separation between the $t = 4$ treatments. Because so few variables have been selected, it is likely that most of the dependence between them is due to the between-treatment variation. The variables in M_2, on the other hand, are each selected independently of one another. That is to say, if there is dependence among the variables within a given treatment, the variable selection method described in Section 3.2 would not take it into account.

7. Simulation study

7.1 *Design*

In both of our variable selection criteria, we assume a degree of conditional independence. For the method described in Section 3.1, we assume that $\boldsymbol{S}_{\boldsymbol{W}}^{-1}\boldsymbol{S}_{\boldsymbol{B}}$ can be estimated reasonably by $\boldsymbol{W}$, and in Section 3.2, we do not make any attempt to account for the within-treatment dependence structure. Therefore, we present the following simulation study to show the effect of these assumptions on the composition of the subsets of selected variables in the presence of two different within-treatment dependence structures, taking $t = 4$ in all cases.

The first within-treatment dependence structure we consider is the intraclass correlation structure,

$$\boldsymbol{S}_{\boldsymbol{W}}^{(1)} = (1-\rho)\boldsymbol{I} + \rho\boldsymbol{J},$$

where $\boldsymbol{J}$ is the $d \times d$ matrix of ones. The second within-treatment dependence structure we consider is the following:

$$\boldsymbol{S}_{\boldsymbol{W}}^{(2)} = \begin{pmatrix} \boldsymbol{S}_{\boldsymbol{W}}^{(1)} & \rho\boldsymbol{J} \\ -\rho\boldsymbol{J} & \boldsymbol{S}_{\boldsymbol{W}}^{(1)} \end{pmatrix},$$

corresponding to two groups of d variables having intraclass correlation within each group and being negatively correlated with the variables in the other group. For each of the two dependence structures, we consider $\rho = 0.0, 0.3, 0.6, 0.9$, corresponding to independence, weak dependence, moderate dependence, and strong dependence, respectively. And we generate binary data matrices $\boldsymbol{X}$ defined by $X_{ij}^k = I(Y_{ij}^k \leq 0)$, where

$$\boldsymbol{Y}_{ij} \sim \mathrm{N}\left(\Phi^{-1}(\boldsymbol{p}_i), \boldsymbol{S}_{\boldsymbol{W}}^{(*)}\right)$$

for $j = 1, \ldots, n_i$, Φ^{-1} is the inverse of the standard normal c.d.f., and $\mathrm{E}(\boldsymbol{X}_{ij}) = \boldsymbol{p}_i$ for $i = 1, \ldots, t$ and $j = 1, \ldots, n_i$.

Two different structures for the set of treatment mean vectors $\boldsymbol{p}_i$, $i = 1, \ldots, t$, or equivalently for $\boldsymbol{S}_{\boldsymbol{B}}$ are also considered. The first structure is a general case, denoted by $\boldsymbol{S}_{\boldsymbol{B}}^{(1)}$, with treatment mean vectors of dimension $d = 250$, where the probabilities p_{ik} take on the values 0.0, 0.10, 0.50,

Table 3
Proportion of 100 simulations that the 19 variables in M_2 are selected by our proposed methods in the case of the simulation study most resembling the Nakatsu data. The three variables in M_1 are indicated by a dot. Methods 1 and 2 refer to Sections 3.1 and 3.2, respectively.

		Method 1				Method 2			
		0.00	0.30	0.60	0.90	0.00	0.30	0.60	0.90
	X^{9}	1.00	1.00	1.00	1.00	0.17	0.15	0.20	0.49
	X^{12}	0.98	0.95	0.99	0.99	0.03	0.05	0.22	0.43
•	X^{13}	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	X^{14}	0.70	0.66	0.66	0.63	0.00	0.00	0.00	0.00
	X^{19}	0.82	0.79	0.89	0.92	0.04	0.00	0.00	0.02
	X^{32}	0.99	1.00	1.00	1.00	0.24	0.40	0.47	0.74
•	X^{34}	1.00	1.00	1.00	1.00	0.72	0.45	0.32	0.58
	X^{36}	0.95	0.95	0.96	0.97	0.05	0.14	0.29	0.51
	X^{39}	0.90	0.81	0.72	0.83	0.03	0.00	0.00	0.04
	X^{40}	0.56	0.60	0.62	0.68	0.02	0.00	0.00	0.00
	X^{43}	0.64	0.57	0.69	0.66	0.04	0.00	0.00	0.01
	X^{45}	0.79	0.66	0.61	0.64	0.03	0.02	0.01	0.00
	X^{46}	0.76	0.76	0.76	0.80	0.04	0.02	0.00	0.07
	X^{48}	0.96	0.96	0.99	0.97	0.34	0.12	0.16	0.28
	X^{49}	0.62	0.56	0.62	0.79	0.09	0.00	0.00	0.02
	X^{53}	0.60	0.63	0.63	0.55	0.01	0.00	0.00	0.02
•	X^{54}	1.00	1.00	1.00	1.00	0.89	0.97	0.94	0.94
	X^{55}	0.65	0.66	0.62	0.70	0.10	0.00	0.00	0.00
	X^{84}	0.97	0.94	0.94	0.98	0.01	0.00	0.00	0.00

and 0.90. The second structure of treatment mean vectors is the situation, denoted by $S_B^{(2)}$, corresponding to the $d = 84$-dimensional sample treatment mean vectors for the Nakatsu et al. (2000) data.

Note that the dimensions of $\boldsymbol{X}$ depend on the within-treatment covariance structure under consideration. For $S_W^{(1)}$, $\boldsymbol{X}$ has dimension $n \times d$, and for $S_W^{(2)}$, $\boldsymbol{X}$ has dimension $n \times 2d$. The dimension of $\boldsymbol{X}$ also depends on the between-treatment covariance structure. For the general case, $S_B^{(1)}$, we take $n_1 = n_2 = n_3 = n_4 = 25$ and $n = 100$, and for the case corresponding to the Nakatsu et al. (2000) data, $S_B^{(2)}$, we take $n_1 = 23$, $n_2 = n_3 = n_4 = 22$, and $n = 89$.

7.2 *Results*

The two different methods of variable selection are applied to 100 simulated data matrices for each of the 16 different cases (i.e., $2S_W^{(*)}$ structures $\times$ $2S_B^{(*)}$ structures $\times$ 4ρ values $= 16$ cases). We observe in general that, as the within-treatment dependence gets stronger, the estimated number of variables selected by the method described in Section 3.1 increases. This indicates that the method does indeed take the within-treatment covariance structure into account in the variable selection methodology even though we use $\boldsymbol{W}$ as an estimate of $S_W^{-1}S_B$. A similar trend is not observed in the results for the variable selection methodology described in Section 3.2. This result is expected because this method ignores dependence between the variables by design.

For both variable selection methods, it was observed that the standard deviation of the number of variables selected increases as the strength of the within-treatment dependence between variables increases. This increase in variation is likely due to the fact that very highly correlated variables will, most likely, either be selected or not selected as a group. This is further supported by the results for the $S_B^{(2)}$, $S_W^{(1)}$ case, where we observe the proportion of the 100 simulated data sets for which each variable in M_2 is selected. For example, variables X^{32} and X^{36} are both selected more often by the first method as the strength of the within-treatment dependence increases.

8. Discussion

In this 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 that require further attention.

First, the variable selection method from Section 3.1 may not be ideal for the particular application of microbial community analysis because it tended to select only a small number of variables, while the method from Section 3.2 selected a more reasonably sized subset of variables for the Nakatsu et al. (2000) data. However, as d gets very large, perhaps the first method would produce more reasonably sized subsets than the second method. Also, due to the design of the first method, 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.

Second, we realize that the assumption of conditional independence may have an effect on our ability to evaluate subsets of dependent variables via cross-validation. It is likely that a classification rule that accounts for dependence between selected variables would classify some observations correctly that were not correctly classified by the conditionally independent classification rules. However, it does not seem likely that many observations classified correctly by the conditionally independent classification rules would be incorrectly classified by a rule that accounted for dependence. In addition, because the goal of the application, microbial community analysis, is not necessarily to select the best predictive subset but rather to select a subset of the observed microorganisms as candidates for further research, the conditionally independent classification rules give a reasonable indication of the validity of the subsets of selected variables. For these reasons, we consider the results of cross-validation using the conditionally independent classification rules to be adequate for this application.

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 as for continuous data, with some modification.

Acknowledgements

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 like to thank Judy Lindell and the entire tillage research group at Purdue for their technical assistance and continual support as well as the associate editor and the two referees for their helpful comments.

Résumé

Pour comprendre l'influence des communautés microbiennes sur la productivité d'une culture, l'identification et la caractérisation des communautés microbiennes de la rhizosphère du sol sont nécessaires. Les profils caractéristiques des communautés microbiennes sont obtenus, à partir d'échantillons de sol, par la méthode de l'électrophorèse en gel de gradient de dénaturation (DGGE) de rADN 16S amplifiés issu d'une PCR (polymerase chain reaction). Ces profils caractéristiques, généralement appelés empreintes génétiques, peuvent être représentés sous la forme de vecteurs binaires multidimensionnels. Nous nous intéressons au problème de la modélisation et de la sélection des variables pour ces données binaires multidimensionnelles et présentons une application de notre méthodologie dans le contexte d'une expérimentation agricole contrôlée.

References

Akkermans, A. D. L., van Elsas, J. D., and de Bruijn, F. J. (1996). *Molecular Microbial Ecology Manual*. Norwell, Massachusetts: Kluwer Academic.

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

ben Omar, N. and Ampe, F. (2000). Microbial community dynamics during production of the Mexican fermented maize dough pozol. *Applied and Environmental Microbiology* **66**, 3664–3673.

Bevivino, A., Sarrocco, S., Dalmastri, C., Tabacchioni, S., Cantale, C., and Chiarini, L. (1998). Characterization of a free-living maize-rhizosphere population of *Burkholderia cepacia*—Effect of seed treatment on disease suppression and growth promotion of maize. *FEMS Microbiology Ecology* **25**, 225–237.

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.

Crookston, R. K., Kurle, J. E., and Lueschen, W. E. (1988). Relative ability of soybean, fallow, and triacontanol to alleviate yield reductions associated with growing corn continuously. *Crop Science* **28**, 145–147.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology* **26**, 297–302.

Dick, W. A. and Van Doren, D. M. (1985). Continuous tillage rotation combinations effects on corn, soybean and oat yields. *Agronomy Journal* **77**, 459–465.

Emrich, L. J. and Piedmonte, M. R. (1991). A method for generating high-dimensional multivariate binary variates. *American Statistician* **45**, 302–304.

Felske, A., Wolterink, A., van Lis, R., de Vos, W. M., and Akkermans, A. D. L. (2000). Response of a soil bacterial community to grassland succession as monitored by 16S rRNA levels of the predominant ribotypes. *Applied and Environmental Microbiology* **66**, 3998–4003.

Griffith, D. R., Kladivko, E. J., Mannering, J. V., West, T. D., and Parsons, S. D. (1988). Long-term tillage and rotation effects on corn growth and yield on high and low organic matter, poorly drained soils. *Agronomy Journal* **80**, 599–605.

Hand, D. J. (1997). *Construction and Assessment of Classification Rules*. New York: Wiley.

Iwamoto, T., Tani, K., Nakamura, K., Suzuki, Y., Kitagawa, M., Eguchi, M., and Nasu, M. (2000). Monitoring impact of *in situ* biostimulation treatment on groundwater bacterial community by DGGE. *FEMS Microbiology Ecology* **32**, 129–141.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bulletin Society Vaud Science National* **44**, 223–270.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edition. New York: Chapman and Hall.

Miethling, R., Wieland, G., Backhaus, H., and Tebbe, C. C. (2000). Variation of microbial rhizosphere communities in response to crop species, soil origin, and inoculation with *Sinorhizobium meliloti* L33. *Microbial Ecology* **40**, 43–56.

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. *Proceedings of the 15th Conference of the International Soil Tillage Research Organization (ISTRO)*. Fort Worth, Texas: ISTRO.

Nuamah, I. F., Qu, Y., and Amini, S. B. (1996). A SAS macro for stepwise correlated binary regression. *Computer Methods and Programs in Biomedicine* **49,** 199–210.

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

Sohn, S. Y. (1999). A comparative study for stepwise correlated binary regression. *Computer Methods and Programs in Biomedicine* **59,** 181–186.

Tannock, G. W., Munro, K., Harmsen, H. J. M., Welling, G. W., Smart, J., and Gopal, P. K. (2000). Analysis of the fecal microflora of human subjects consuming a probiotic product containing *Lactobacillus rhamnosus* DR20. *Applied and Environmental Microbiology* **66,** 2568–2588.

Torsvik, V., Sorheim, R., and Goksoyr, J. (1996). Total bacterial diversity in soil and sediment communities—A review. *Journal of Industrial Microbiology* **17,** 170–178.

Torsvik, V., Daae, F. L., Sandaa, R. A., and Øvreås, L. (1998). Novel techniques for analysing microbial diversity in natural and perturbed environments. *Journal of Biotechnology* **64,** 53–62.

Turco, R. F., Bischoff, M., Breakwell, D. P., and Griffith, D. R. (1990). Contribution of soil-borne bacteria to the rotation effect in corn. *Plant Soil* **122,** 115–120.

van Hannen, E. J., Mooij, W., van Agterveld, M. P., Gons, H. J., and Laanbroek, H. J. (1999a). Detritus-dependent development of the microbial community in an experimental system: Qualitative analysis by denaturing gradient gel electrophoresis. *Applied and Environmental Microbiology* **65,** 2478–2484.

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

Venables, W. N. and Ripley, B. D. (1999). *Modern Applied Statistics with S-Plus*, 3rd edition. New York: Springer.

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: Examples and Methods for p-Value Adjustment*. New York: Wiley.

Received April 2001. *Revised December* 2001.
Accepted December 2001.