Build Predictive Models: Issues and Applications

Hongjie Wang [1]

February 4, 2002

[1]This paper is dedicated to my dear wife Jessica Q. Wang, whose unconditional love and support continue to motivate me, to my advisor Dr. Hanif D. Sherali whose teaching inspires me to pursuit excellence in everything I do, to Raye Shorten who taught me Statistics and to all my esteemed colleagues in Circuit City who I had the privilege to work with in the past one year (2001).

Contents

1 Introduction

In this paper, we list some important statistical issues and techniques in model building from a practitioner's stand point. The emphasis is on presenting useful heuristics from data mining and machine learning and their relevant application in predictive modeling. Whenever appropriate, outlines of the mathematical derivations are given to help understand the techniques. Illustrative examples using real and artificial data are used. Most of the techniques are implemented in SAS programs.

2 Significance Test

2.1 Data-driven vs. Model-driven

Data-driven approach almost inevitably leads to the concept of a model even if that model is only vaguely specified. The difference is that in a data-driven approach models are chosen to satisfy conditions imposed by or suggested by the data or our knowledge about that they represent, whereas in the model-driven approach the data are either assumed to fit a pre-specified model, or if they do not, they are adapted to fit it by procedures like transformations. Experienced statisticians should use both approaches. Often time, people consider the nonparametric or distribution-free methods as data driven too. Chi- Square test is a nonparametric test, so is the Spearman's rank correlation. But one of the first true data driven test is Fisher's exact permutation test.

Suppose we have subject two groups to two different treatment and measures the outcomes. The data looks like the following:

Group A receiving treatment A (0, 3.5, 4.2)

Group B receiving treatment B (4, 4.6, 7, 9, 10.5).

The question is whether the treatment has a significiant effect.

Thus, H_0:*There is no differential response in the measured characteristic between treatment* and the alternative hypothesis H_1 is *The measured characteristic is greater for treatment B.*

A commonly used test is the t-test (one tail). Notice t-test assumes that both population are distributed normal with each variance and possibly different means. Under such strict assumption, t-test is optimal in detecting location shifting. However, as we know, such assumptions may not be true or

difficult to verify. Randomization (permutation) test can be used in this case. We use the above example to illustrate the concept. If group C is the set of all values in group A and B, then groups A and B are discordant if group A is unlikely group among all groups of that size that could be obtained by random sampling from group C. Notice if group A is discordant then group B is also discordant. Discordance suggests rejection of H_0 is appropriate. Under randomization, all possible permutations of 3 from 8 units are equally likely, but anlikely group is one more likely under H_1 than H_0. It is trivial to show that there are totally $56(C_3^8)$ distinct sets of 3 items that may be selected from 8 different items. If H_1 is true, then group A values will tend to be lower than group B values. One can use the mean or the sum to compare among groups. We can tell from visual inspection that the smallest three groups are $(0, 3.5, 4), (0, 3.5, 4.2), (0, 3.5, 4.6)$. Thus, if we proceed the randomization, there are only $2/56 = 0.0357$ chance that we obtain a group with smaller sum that group A. Thus, we reject H_0 at 3.57% level. It should be clear now why it is called permutation test. Furthermore, it is data-driven and does not assume any distribution properties. Fisher is among the first to design such tests. It is computationally very intensive and therefore was not feasible in the early days. However, these are exact tests with exact P-values while the traditional t-test is an approximation. It has been shown that when the distribution assumption does hold, the t-test result is very close to exact permutation test.

One of the more familiar exact test is the Fisher's exact test in contingency table analysis. It is used widely to detect association when the cell sizes are small. People has the misunderstanding that Chi Square is better when the cell sizes are large enough. In fact, the exact test should always be better. However, it is computationall infeasible when the cell sizes are large.

Peter Sprent lists exact permuation tests for most of the traditional parametric tests in his book *Data Driven Statistical Methods.*

The relationship between the traditional tests and their permutation counterparts can be better understood by examining the relationship between Fisher's linear discriminant analysis and logistic regression. The formal is model based (one can tell from the fact that all the coefficients of the models are determined by the sufficient statisitcs, mean and covariance). Logistic regression is semi-parametric and its estimation relies up Maximum-Likelihood method.

2.2 A Note on Statistical Power

Two kinds of "wrong" decision may result from a hypothesis tes. Rejecting H_0 when it is true is called Type I error. For example, we know that stepwise regression tends to increase Type I error since it tends to bring in unimportant variables. (The H_0 in that application is $\theta_i = 0$.) In an exact test, the probability of committing such an error is $\alpha 100\%$ where α is the p value wer report. Accepting H_0 when H_1 is true is an error of Type II. The probabiliy of making this mistake is β (and β is not $1-\alpha$!). We generally do not know β and its values depends on variance and sample size and the choice of α. The quality $1-\beta$ is called power. The power equals the probability of getting a result in the chosen region (thus rejecting H_0) when H_1 is indeed true. When design a test, one usually set the power (to 0.8 for example) and the sample size can be determined as a result. It should be intuitive that a larger sample size corresponds to higher power, as it is more likely to bring the effect to surface. Also, a smaller α in effect makes the test more powerful since it makes easier to reject H_0 when H_1 is true. However, if a test is conducted and H_0 is rejected at a specified level, it is inappropriate at this point to claim that the test has a higher power. Such after-fact power calcuation is confirming a self-fulfiling promise.

Another related concept is the efficiency of the test. It is measured by the ratio of the sample size needed to attain the same power for detecting a given true parameter included in H_1 for a chosen α.

We use an interesting example to illustrate the importance of power of significance tests.

	Promoted	Not Promoted	Total
Group A	1	31	32
Group B	10	58	68

	Promoted	Not Promoted	Total
Group A	2	46	48
Group B	9	43	52

The objective is to show if there is an association between the group membership and the promotion. It looks like group B is favored since it has a higher promotion rate. There are totally 11 being promoted in both cases. If there is no assoication between group membership and promotion, then we would expect to see the group A's promotion rate is also 11 %.

In the first table, the promotion rate of group A is $1/32 = 3.125\%$ and the promotion rate of group A in the second table is $2/48 = 4.167\%$. Both of them are compared agaist the expected response rate of 11%. Surprisingly enough, using a Fisher's exact test, we reject association in the first table and fail to reject the association of the data in the second table.

Here is the SAS program conducting this exercise.

```
data data1;
  input response treatment count;
  cards;
1 1 1
1 0 10
0 1 31
0 0 58
  ;
run;

data data1(keep=response treatment);
   set data1;
   do i=1 to count;
      output;
   end;
run;


proc freq data=data1;
     table response*treatment /exact;
run;
```

```
                Fisher's Exact Test

        Cell (1,1) Frequency (F)        58
        Left-sided Pr <= F          0.0765
        Right-sided Pr >= F         0.9892

        Table Probability (P)       0.0657
```

```
Two-sided Pr <= P            0.1001

        Sample Size = 100


       Fisher's Exact Test

Cell (1,1) Frequency (F)        43
Left-sided Pr <= F          0.0351
Right-sided Pr >= F         0.9942

Table Probability (P)       0.0293
Two-sided Pr <= P           0.0534

        Sample Size = 100
```

This is certainly surprising since the second table shows a relatively higher promotion rate of group A. The problem is not the test itself. Rather, the power of the test is in question. This anomaly may partly be accounted for by the fact that there are relatively fewer (32) of the group A in the first table.

2.3 Boostrap

It has been shown mathematically and empirically that a random sample reflects the characterisic of the population fairly well. Given a sample $\{x_1, x_2, \ldots, x_n\}$, we create the order statistics $\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$ where $x_{(i)} \leq x_{(i+1)}$. The sample cdf $S(x)$ can be defined as a stepwise function. For sample with decent size 50, the difference $F(x_i) - S(x_i)$ is usually less than 0.2. The technique of boostrap uses the idea of usings samples of samples to further explore characteriscs that reflect those in the original sample and so in turn may relect those in the population. Although boostrapping can be justified as a valid technique by complicated mathematics, it is not exact test. However, usually, it is more accurate and informative than those obtained by fitting a wrong model to data by assuming normality when it is not. The central step of boostrapping is the sample with replacement. Instead of going

over the tedious mathematics, we show two examples of boostrapping, one estimating the mean, the other cross validation.

3 A Note on Outliers

Outliers detection is as complicated a problem as variable selection. To begin with, a precise definition is not possible. In broad teams for univariate data an outlier is an observation so remote from other observtions as to cause surprise. Surprise is, by its nature, subjective. If a distribution is known, then some parametric tests can be conducted to detect outliers. It is not practical in observational study, where usually the distribution of the variable is not known. In the context of multivariate analysis, outlier is even more difficult to qualify. An observation that looks strange in one dimension may be perfectly normal in the multidimensional space. In dependency anlaysis such as regression, usally the foucs is to find points that some how does not fit an otherwise well defined linear or nonlinear relation.

There are two widely used techniques in dealing with distribution free univariate outlier detection. These heuristics turn to be useful in multivariate analysis as well uisng as univariate pre-processing and screening.

- Trimmed Mean: The observations are arranged in ascending order and then, we delete the top $t\%$ and the bottom $t\%$ of the observations. For sample with large enough size, t can be taken anything from 10 to 20.
- Windsorization: We shrunk the extreme observations to the value of the remaining observations of greatest magnitude in each tail and thus reducing their influence. The reason for this precedure is that evey often an outlier contains useful information and there is not justification to just simply delete it. This technique is very useful in the regression where we may see a linear relationship levels off after the predictor goes beyond certain range.

A practical outlier detection procedure is *median absolute deviation*. An observation x_0 is considered an outlier if

$$\frac{|x_0 - med(x)|}{med[|x - med(x)|]} > 5. \tag{1}$$

Consider the following data $\{-3, 11, 21, 24, 25, 25, 27, 29, 31, 44, 57\}$. The median is 26 and the absolute median deviations are $\{29, 15, 5, 2, 1, 1, 1, 3, 5, 6, 18, 31\}$. The median of this sequence is 5. Thus, using the above formula we would classify -3 and 57 as outliers.

4 Statistical Issues in Model Specification

Technically speaking, model specification is population level problem while model estimation is a sample level problem. We shall not make such fine differentiation here, as quite often, they are related. For example, one may have specified a model $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots$ (We assume that both x_1, x_2 are relevant dimentions in the population level joint distribution.) Suppose we have collected some data from this population. Due to the sample error, the optimization algorithm finds that the estimate of β_1 is very close to 0. As a result, we drop this variable from our model. Then, the resulting model consititute a biased representation between y and x at the population level. In the next several subsections, we will discuss this topic in detail. But first, we use one example to illustrate a framework on how model specification tests can be carried out.

Let Ω denote some information set and we specify the linear regression model

$$y = \alpha + \beta x + \mu. \tag{2}$$

The model is said to be misspecified if $E(y|\Omega) \neq \beta x$. Suppose that z is another independent variable that also belongs to Ω, then we can have another alternative model

$$y = \alpha + \beta x + \lambda z + \mu. \tag{3}$$

The key to establish the difference between these two models is the t statistics for $\lambda = 0$, which may be denoted as t_λ. The distribution of this statistics depends on the the stochastic process that generated y.

Similarly, we can test the nonlinearity condition by looking at the alternative model

$$y = \alpha + \beta(1 + (x^\lambda - 1)/\lambda) + \mu. \tag{4}$$

In this case, we need to test for $c = 0$ in the following equation

$$y - \hat{\alpha} - \hat{\beta} x = a + bx + c(xlog(x) - x + 1) + residual. \tag{5}$$

Notice, the alternative model uses Box-Cox transformation. For a detail account of various model specification tests, see *Model Specification Tests and Artificial Regressions* by MacKinnon.

4.1 Population Selection Bias

We often have to deal with this problem. T. M. Smith, the President of Royal Statistical Society, in his address titied "Populations and Selection: Limitations of Statistics", mentioned selection bias is one of the most important and difficult issues statisticians face. One of the most favourite examples is the aircraft survival study done by Abraham Wald. The data available were the patterns of bullet holes in those aircrafts returning to base, and the problem was where to place additional armour. Wald's solution was to put the armour where there were no holes, the argument being that that was where the planes that did not return must have been hit.

We often use data from previous campaigns to develop predictive modles. Ideally, the campaign should have used randomly selected customers. However, this ususally is not the case. Bias therefore will be introduced into the models. (Response models we built for Circuit City is a case of example.)

In linear regression, sometimes, we want to estimate the population multiple correlation coefficnet from a sample. It is long known that the values derived by the sample tend to be "deceptively" large. Lots of research has been devoted to various adjustments.

Most of the classic statistical inference techniques were designed to handle *experimental design data* instead of *observational data*. In real like, it is almost certain that we would encounter high level dependency among variables, and omission of important variables. Such difficulties can be largely overcome by designed experiments, balance and randomization. As Cox and Snell point out in *The Choice of Variables in Observational Studies*, randomization of an omitted important variable leads to an increased error variance and to a seriously incomplete understanding, but not to a "biased" conclusion. In addition, taking some extra appropriately chosen observations can reduce the non-orthogonalities in the data. But such problems and their resulting biases are likely to remain in an observational study.

One of the basic assumptions of the multiple regression model is that the values of the independent variables are known constants and are fixed by the researcher before the experiements. Only the dependent variable is free to vary from sample to sample. This is called fixed linear regression models.

Such models are hardly realistic in practice when we have to control over the independent variables at all. An alternative model allows predictors to be random variables as well and is called random model. Although the β coefficients from ML are the same from both models, the distributions are very different. In fact, most of the traditional inferential statisitcs we use assume the fixed model where the estimate of β is normally distributed. The potential problem of this practice is that the random error introduced from the sample data tend to be capitalized in the optimization process. Yin and Fan (2001) gives a detailed account on the subject of R^2 adjustment in regression models in light of the bias.

Another topic that is related to population selection is population shifts. One of the implicit assumption that is intrinsic to the statistical paradigm is that populations are static. They do not change over time. This obviously is not realistic. How fast and drastic a population will change depends on the industry. When we deal with a siutation where we know the characteristic of the population changes very frequently, we have to double our effort to build more parsimious models so that we do not capitalize any noise or abnormaly that is unique to the data under study. It is also important to point out that not every population shift has an impact. For example, in a classification problem, the shift of class prior $p(y)$ and distributions of the classes $p(x|y)$ may not have any impact on the classification rules, so long as the changes of x affect the classes in the same way. On the other hand, if $p(y|x)$ changes, then most likely, the model has to be refitted. See detail on this subject, we refer to *The Impact of Changing Populations on Classifier Performance* by Kelly, Hand and Adams.

4.2 Error and Bias Conflicit

Let y be a continuous target variable with an unknown pdf $p(y)$. Let $\hat{p}(y)$ be the estimated pdf from the data. (Keep in mind that the essence of all modeling is to estimate $p(y)$.) The difference between $p(y)$ and $\hat{p(y)}$ can be measured by the following Kullback-Leibler operator

$$K(p, \hat{p}) = \int dy p(y) log[\frac{p(y)}{\hat{p(y)}}]. \tag{6}$$

It can be shown that the expected value of the above equation can be decomposed into two parts: variance and bias. The bias is supposed to measure

how closely the algorithm 's average guess matches the target and the variance how much the algorithm's guess bounces around for different datasets. Modifications on the model (algorithm) tend to have an opposite effect on the bias and the variance. This observation is based on the well known statistical property that an increase in the number of degrees of freedom usually leads to a smaller bias and a higher variance. The Neural network is a perfect example. NN is very flexbile and does not require any assumption on the data. It is proven to be a universal approximator. As a result, the functional form it derives is likely to have very small bias. On the other hand, it requires lots of parameters and tend to suffer from overfitting problem, which is a consequence of large variance.

Let X be a n vector and $y(X)$ is a random scalar output. We assume that a sample of N observations $D_N = \{X^k, y^k = y(X^k)\}_{k=1}^N$ and there exists an unknown regression function μ such that $y(X^k) = E(y(X^k)] + \epsilon^k = \mu(X^k) + \epsilon^k$. Let's further assume that μ has expected value of 0 and variance δ. Regression problem is to find a parameterized function $f(X, \theta, D_N)$ that is a good approximation of $\mu(X)$. Consider the following:

$$E[((y(X) - f(X))^2] = \tag{7}$$
$$E[(y(x) - \mu(x))^2] + E[(f(X) - \mu(X))^2] = \tag{8}$$
$$\delta^2 + (E[f(x)] - \mu(x))^2 + E[(f(X) - E[f(X)])^2]. \tag{9}$$

Notice the first part is the error and the second part is variance.

4.3 Inclusion of Irrelevant Variables

We concentrate on the sample case of regression model with 2 variables. The mathematics for multiple variable cases are more complicated. The ideas, however, are similar.

Let $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ be the true model for the population. Under least square condition, we can derive the standard error of b_1, b_2 as

$$s_1 = \sqrt{\frac{||y - \hat{y}||}{||x_1 - \bar{x_1}||(1 - r^2_{x_1,x_2})(n-3)}}, \tag{10}$$

$$s_2 = \sqrt{\frac{||y - \hat{y}||}{||x_2 - \bar{x_2}||(1 - r^2_{x_1,x_2})(n-3)}} \tag{11}$$

where $r_{x1,x2}$ is the correlation of x_1, x_2.

Suppose the true model is $y = \alpha + \beta_1 x_1 + \epsilon$, but a model of $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ is constructed. In other words, we have included the irrelevant variable x_2 in the model. What is the consequence of much a misspecifiction?

It can be shown (see Rao and Miller 1971) that $E(b_1) = \beta_1$ and $E(b_2) = 0$. Thus, the coefficients are still unbiased. With any luck, we may find b_2 to be nonsignificant and drop x_2 from the model as a result. But, such inclusion has impact on the standard error of b_1. From the formula listed above, it follows that the standard error of b_1 will be inflated roughly by $\frac{1}{\sqrt{1-r^2_{x_1,x_2}}}$. We will see that such inflation is really due to the collinearity and it happens among revelant variables too. But, since x_2 is not relevant to begin with, paying such a price is not worthwhile.

We illustrate the point by looking at the following SAS program:

```
%let n=10000; %let nvar=7;


proc iml;
R={1    0.72     0.63  0.54 0.45 0.3  0.2,
   0.72    1     0.56  0.48 0.4  0.2  0.3,
   0.63 0.56        1  0.42 0.35 0.1  0.45,
   0.54 0.48     0.42     1 0.3  0.3  0.2,
   0.45 0.4      0.35  0.3    1  0.12 0.34,
   0.3  0.2      0.1   0.3  0.12    1 0,
   0.2  0.3      0.45  0.2  0.34    0    1 };
means={0 0 0 0 0 0 0};
sts={1 1 1 1 1 1 1};
 Z=shape(0,&n,&nvar);
 do n=1 to &n;
   do j=1 to &nvar;
     Z[n,j]=normal(0);
   end;
 end;
 call eigen(L,V,R); L=diag(L);
 X=(Z*sqrt(L)*V');
do i=1 to &n;
   do j=1 to &nvar;
```

```
      X[i,j]=X[i,j]*sts[j]+means[j];
   end;
end;
 varnames='x1':"x&nvar";
 create xdata from X[colname=varnames];
 append from X;
 close xdata;
 call eigen(L,V,R); L=diag(L);
 V=V*sqrt(L);
quit;


data xdata;
   set xdata;
   y11=x6+rannor(10);
   y12=x1+rannor(10);
   y21=x6+3*x7+rannor(10);
   y22=x1+3*x2+rannor(10);
run;
data sample;
   set xdata;
   if ranuni(100)<=0.2;
run;

proc standard data=sample m=0 std=1;
   var y11 y12 y21 y22 x1 x2 x6 x7;
run;

/* Including Irrelevant Variable
   1. true model y11=x6,
      fit model y11=x6 x7, x6 and x7 uncorrelated
   2. true model y12=x1,
      fit model y12=x1 x2, x1 and x2 correlated
*/

proc reg data=sample;
   model y11=x6 /noint;
   model y11=x6 x7 /noint;
```

```
   model y12=x1 /noint;
   model y12=x1 x2 /noint;
run;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
x6	1	0.99467	0.02218	44.84	<.0001

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
x6	1	0.99409	0.02219	44.79	<.0001
x7	1	-0.01870	0.02210	-0.85	0.3975

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
x1	1	0.98882	0.02276	43.44	<.0001

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|

x1	1	0.96012	0.03248	29.56	<.0001
x2	1	0.04031	0.03255	1.24	0.2157

In the first case, x_6, x_7 are not correlated. Adding the extra x_7 does not have any impact on x_6. In the second case, by adding x_2, we see that the standard error of x_1 goes up.

4.4 Omissions of Important Variables

Avery *el al.* examine the impact of not including important variables on the credit scores from major credit bureaus. Typically, variables reflective of local economic conditions are not included in scoring. A person who otherwise is very credit worthy may have been forced to default due to a local economic situation. The resulting low score is only predictive if the same economic siutation repeats itself in the future. Marais and Wecker (1998) examine the impact of lead on IQ. They argue that the conventional belief of negative impact of lead on IQ is overstated by not including parents's IQ in the regression equation. Suppose we have $IQ = f(lead, Parents' IQ, Others)$. It is conceivable that IQ is positively correlated with parents' IQ and parents' IQ is negatively correlated with the exposure to lead. Thus, by leaving out the parents' IQ in regression, the coefficient of lead is likely to be negatively biased, that is, it is more negative than it should be. The bias introduced by omitting important variables are well documented in various industries where predictive models are also used to descriptive purposes.

In making causal inferences from statistical analysis, one of the most commonly asked question is "Have you counted for factor B?". Factor B could be a confounding factor which is not included in the model and whose inclusion may alter the relationship between the target variable and the variables already in the model. In this section, we give a brief mathematical outline on this subject. Specifically, we will explore the impact of omitting variables on the coefficient estimates of the variables that are in the model. Once again, we will restrict our discussion on regression models with no more than 2 predictors for simplicity.

Suppose $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ is the true model, but for various reasons, x_2 is included from the equation. If x_1 and x_2 are not correlated, the case

is relatively simple. In this situation, the predictors are not fully explaining the variances in y and the model encountered too much heterogeneity. The consequence this the inflated error term ϵ. This causes the standard error of b_1 to go up. However, the inflated error term in this case is not related to x_1. Thus, b_1 is still unbiased.

When x_1 is correlated with x_2, then by skipping x_2, we are essentially fitting the model $y = \alpha + \beta_1 + \epsilon'$ where $\epsilon' = \epsilon + \beta_2 x_2$. Now, it should be clear that one of the fundemental assumptions of least square regression model is voilated, namely, predictors are uncorrelated with ϵ. Since x_1, x_2 are correlated, x_1 is therefore correlated with the new error term ϵ'. Under this condition, it can be shown (see Rao and Miller, 1971) that $E(b_1) = \beta_1 + \beta_2 r_{x_1,x_2}$. The estimate in this case is biased! Intuitively, when a relevant variable is exluded from the estimation of a regression model, the variables left will pick up some of the impact of the excluded variable on y. The dirction of the bias (positive or negative) will depend on both the relationship between the omitted variable with y and with the variables left. For example, if β_2 is positive and r_{x_1,x_2} is positive, then b_1 will tend to be larger than it should be.

The impact of omitting relevant variables on the standard error of the variables left in the model is somewhat surprising. They actually tend to get smaller. This is due to the fact that removing variables always reduce collinearity. We use the following SAS problem to illustrate.

```
/* Exclude Relevant Variable
   1. True  model: y21=x6 x7;
      fit model : y21=x6, x6 x7 uncorrelated;
   2. True model: y22=x1 x2;
      fit model : y22=x1, x1 x2 correlated;
*/

proc reg data=sample;
   model y21=x6 x7 /noint;
   model y21=x6 /noint;
   model y22=x1 x2 /noint;
   model y22=x2 /noint;
run;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
x6	1	0.99460	0.02251	44.19	<.0001
x7	1	3.02701	0.02241	135.05	<.0001

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
x6	1	0.90121	0.07169	12.57	<.0001

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
x1	1	1.02142	0.03255	31.38	<.0001
x2	1	2.97162	0.03261	91.12	<.0001

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|

x2	1	3.70190	0.02793	132.54	<.0001

In this section, we have discussed the problem of model specification error and its consequences on the coefficient estimates. Specifically, we show the problems with including irrelavant variables as well as omitting important variables. It should be pointed out that we based our analysis on the assumption that we know the "true model", which is never realistic in any observational study (such as database marketing models). In such models, our decisions to detect irrelevant varaibles are most likely driven by the pursuit of a more parsimonious model or picking more meaningful variables our of a pool of competing variables. On the other hand, our ability to include all the important variables is limited by the data collected and the time and resource requirements to sift through thousands of potential variables. As we see from the examples, including irrelavant variables is not as big a problem as omitting important ones. Statisitical techniques are a reasonably helpful in detecting irrelevat variables, since the expected value of their partial slope coefficients is 0. However, due to the standard error and other possible noises in the data, it is possible that for the sample we used for modeling, the estimates is not 0. The issue is further complicated by collinearity. In most cases, however, the estimates will be small with respect to its standard error and a t test should be able to reject the inclusion of such variables. Another indication of irrelevancy is the negligible reduction of R^2 or sometimes increase of adjusted R^2 after the exclusion. The problem of omitting important variables typically results in very small R^2. But other than that, statistical methods, to a substantial degree, are helpless. The best guide to include or exlude variables is again based on the understanding of the meaning of the data and the business domain knowledge.

4.5 Multicollinearity

4.5.1 Definition

Multicollinearity is one of the most common problems one may encounter in building a statisitical model. This is especially true in observational study. In the model, a multicollinearity is said to be present if there is at least one

nonzero constaint a_i such that $\sum_{j=1}^{p} a_j x_j \cong 0$. The ambiguity contained in the defitition is deliberate since multicollinearity almost always exist in the data with various degrees and different impacts.

It is instrumental to make the distinguish between population multicollinearity and sample multicollinearity. Population multicollinearity is present because of the way the independent variables are defined or are related in the population. As a result, every random sample taken from this population will reflect such a relationship. In the observational study, we usually assume sample multicollinearity which has the proptery that relationships among variables in the sample are potentially spurious and possibly change from sampel to sample.

We like to emphasize one important point before we get into the detail of this subject. Despite the problems associated with multicollinearity, the coefficient estimates are still BLUE (best linear unbiased estimator). Thus, even with high multicollinearity, OLS coefficient estimators have minimum variance among the class of unbiased estimators. Unfortunately, with high multicollinearity, "minimum variance" does not mean low variance. We illustrate by an example. Consider the following SAS program. As you can see, we deliberately introduced lots of collinearity among variables. We take a sample from the population and estimate the coefficients. Due to the collinearity problem, the estimates are not reliable. Sometimes, the sign is not correct. However, if one does it 100 times, overall, the expected etimates is very close to the true population parameter!

```
%let nvar=4;
%let n=50000;
  proc iml;

R={1.0   0.5  0.75  0.2,
   0.5   1.0  0.87   0.1,
   0.75 0.87 1.0    0.1,
    0.2  0.1  0.1    1.0};



   means={0 0 0 0};
   sts={1 1 1 1};
```

```
    Z=shape(0,&n,&nvar);
    do n=1 to &n;
      do j=1 to &nvar;
        Z[n,j]=normal(0);
      end;
    end;
    call eigen(L,V,R); L=diag(L);
    X=(Z*sqrt(L)*V');
   do i=1 to &n;
      do j=1 to &nvar;
         X[i,j]=X[i,j]*sts[j]+means[j];
      end;
   end;
    varnames='x1':"x&nvar"; print varnames;
    create xdata from X[colname=varnames];
    append from X;
    close xdata;
quit;
data xdata;
   set xdata;
   x5=0.4*x1-2.4*x2+1.6*x4+rannor(10);
   y=0.17*x1+0.23*x2+0.25*x3-0.2*x4+0.3*x5+rannor(10);
run;


%macro multest;
data coefslist;
   x3=.;
   x1=.;
   x2=.;
   x4=.;
   x5=.;
run;
%do i=1 %to 100;
   data sample;
      set xdata;
      if ranuni(&i)<=0.005;
   run;
```

```
   proc reg data=sample outtest=coefs(keep=x:);
      model y=x1 x2 x3 x4 x5;
   run;
   data coefslist;
      set coefslist coefs;
   run;
%end;

%mend;

%multest;

proc means data=coefslist;
run;
```

Variable	N	Mean	Std Dev	Minimum	Maximum
x3	100	0.2484554	0.1991863	-0.2912311	0.7532187
x1	100	0.1500954	0.1249968	-0.1522540	0.5604200
x2	100	0.2420948	0.1990741	-0.2009992	0.8834404
x4	100	-0.2098787	0.1298823	-0.5233754	0.1590456
x5	100	0.3075782	0.0671917	0.1309769	0.4714251

4.5.2 Mathematical Definition of Multicollinearity

There is no precise way to measure the severity or degree of multicollinearity. The following rules are often useful.

- Extreme pairewise correlation between two regressor variables; $|r|_{max} = \max_{i,j} |C_{ij}|$ where C is the correlation matrix.
- Small determinant of the correlation matrix.

- One or more small eigenvalues of the correlation matrix. If there is one zero eigenvalue, then an exact linear dependency exists.

- Large variance inflation factors, $VIF(j)$ which are the diagonal elements of the inverse of the correlation matrix. Under the normal linear regression setting $y = x\beta + \epsilon$, the OLS estimators for β has the distribution $b \sim N(\beta, \theta^2(X'X)^{-1})$. Thus, when $X'X^{-1}$ is small, the variance of b is going to be very large. In the perfect situation, when we have orthogonal data, $X'X = I$ and the variance of b is $\theta^2 I$. Usually, when a inflation factor is more than 10, one should look for possible collinearity.

- Large R_j^2, where x_j is predicted using the remaining regressor variables and $R_j^2 = 1 - 1/VIF(j)$. $1 - R_j^2 = 1/VIF(j)$ is called tolerance. A rule of thumb is that a tolerance value less than 0.1 may indicate the presence of multicollinearity.

- High Condition index $\eta_j = \sqrt{\frac{\lambda_{max}}{lambda_j}}$ where $lambda$ are the eigenvalues of the correlation matrix. This mesure was first used by Belsley, Kuh and Welsch (1980). They suggest that condiction indices greater than 30 indicate moderate to strong multicollinearity.;

All the above measures are related mathematically. Steward (1987) gives an detailed mathematical derivation and pointed mentioned that all the methods above are neither necessary nor sufficient. One may also find that some approaches are better than others depending on the data and the results are not necessarily consistent either. For example, we may have situation where there is no large bivariate correlation and yet the level of collinearity is very high. Also, it is important to realize that collinearity can not be viewed in isolation. Its impact on a statistical model can be evaluated only in conjunction with other factors of sample size, R^2, and the magnitude of the coefficients. For example, Mason and Perreault (1991) show that bivariate correlations as high as 0.95 has virtually no effect on the ability to recover true coefficients and to draw the correct inference if the sample size is at least 250 and R^2 at least 0.75.

4.5.3 Indication of the Presence of Multicollinearity

Multicollinearity manifests itself in lots of different ways. Parameter estimates fluctuate dramatically with negligible changes in the sample. Parame-

ters estimates turn out to have wrong sign in terms of theoretical considerations. Theoretically important variables are not significant. Multicollinearity should be suspected also when none of the t-ratios for the regression coefficients are significant and yet the overall F test is very significant. When high multicollinearity is present, switching samples, changing the indicator used to measure a variable in the regression model, or deleting or adding a variable to the equation can all lead to dramatic changes in the size of coefficient estimates. One common cause of all the problems listed above is the large variance of the coefficient estiamtes. We know that multicollinearity leads to large variance of b. However, large variance of b could be caused by things other than the multicollinearity. For example, small sample size, variables with small variances can all lead to b with large variance. What is more unique of multicollinearity is the large covariances between coefficient estimators. In general, the larger the correlation among the predictors, the larger the correlations among the coefficient estimators. For example, in a regression model with two variables x_1, x_2, the correlation between b_1 and b_2 is $-r_{x_1,x_2}$. Thus, if x_1, x_2 are positively correlated, then b_1, b_2 are negatively correlated. If b_1 is less than β_1, then b_2 is likely to overestimate β_2. In any event, when predictors are related, the OLS procedure has difficulty seperating their individual effects on y. This is quite plausible. Partial slope coefficients represents the effect of one predictor on y with all other variables held constant. But when variables x_1, x_2 are correlated highly, then it is virutally impossible to hold x_2 constant when we change x_1. Thus, the interpretations of b_1 and b_2 are shaky.

4.5.4 Dealing with Multicollinearity

Dealing with multicollinearity in observational study is a difficult subject. We list some commonly used techniques along with the possible problems associated with them.

- Nonparametric methods, such as PCA regression, Ridge Regression and Root Regression: Most of these approaches will produce biased estimates, trading for the smaller error term. They are not very practical when the models are supposed to be descriptive as we have to worry about the interpretation of the factors that are derived from the orignal variables.

- Transformation of variables by differencing: This is very commonly used in time-series analysis. For example, if we have account balance of last year and this year as two predictors in the regression model, it is likely that they might be correlated to cause collinearity. In this case, we may perform the transformation of defining $log(x_t) - log(x_{t-1})$.

- Creating ratio or summary variables: This technique can be very effective if the resulting composite variable has clear interpretation.

- Dropping redundant variables: This is the most commonly used approach in dealing with multicollinearity. Such approach, however, should be used with caution. First, unless the true coefficient of the dropped variable is 0, then technically, the model will be misspecified. In general, the consequences of model misspecification- biased coefficient estimators - are more serious than those of the multicollinearity. Secondly, one still has to deal with the problem of which variable to drop. This is not at all obvious. One of the rules used quite often is to select the the variable that loaded the hightest on the component that corresponds to the largest collinear index.

4.6 Interaction

Discuss how one can use Tree based model to detect interaction. Given the sequential or hierarchical manner in which a tree model is built, it naturally contains information regarding possible interaction among variables.

4.7 Specification Uncertainty

Most of the classic statistical procedures are designed for confirmatory analysis where a functional form is theoratically determined. In reality, we have to deal with specification uncertainty. Usually, we estimate a parameter θ such that $p(\theta|X, y)$ where X, y are observed data. Actually, there is another random variable we need to count for, the model.

$$p(\theta|X, y) = \int p(\theta|X, y, M)p(M|X, y)dM. \tag{12}$$

The above expression gives the essence of model averaging. Various cross validation procedures (including holdout and boostraping) are special cases of

model averaging. For a good disucssion on this subject, see *Assessment and Propagation of Uncertainty* by D. Draper in *Journal of the Royal Statistical Society.*

5 Bivariate Association

When we have lots of predictors, one commonly used heuristic in pre-processing is the assessment of bivariate association, that is, determine the relationship between the target variable with each predictor individually. It is important to point out that such a myopic approach is neither necessary nor sufficient for the global variable selection problem. If a variable x_1 is very strongly related to y, that does not automatically mean that x_1 is an important variable in the final model, since we have not evaluated x_1 in the presence of other variables. On the other hand, if x_1 is not important as a stand alone variable to predict y, discarding it is technically not appropriate, since it may become important in the presence of other variables. Indeed, the goal of model selection is in obtaining the subset of variables that will result in the "best model" and not in obtaining the subset that includes variables individually most correlated to the target variable. Despite the problems associated with this greedy-method based approach (we discuss why it is not appropriate later), it is usally a practical starting point and tend to give us lots of information.

5.1 Bivariate Screening is Not Appropriate

A bivariable analysis defines the relationship between one independent variable and one dependent variable. It is widely used as a method to initially select independent variables to be used in multivariable analysis. That is, if the statistical p value of a predictor is greater than a desirable arbitrary value pre-specified, then this predictor will not be allowed to compete for inclusion in the multivariable analysis. However, because this method cannot properly control for possible confounding, using this method to screen variables for subsequent analysis introduces a source of possible error and may cause the rejection or inclusion of inappropriate variables in the multivariable analysis.

Consider the following example

Physical Active (active=1)	Gender (male=1 female=0)	Response (yes=1)	Count
0	1	1	84
0	1	0	1200
0	0	1	80
0	0	0	800
1	1	1	10
1	1	0	800
1	0	1	33
1	0	0	1200

We have two variales, gender and physical activity. If we would ignore physical activitity, we have the following data

Gender (male=1 female=0)	Response (yes=1)	Count
0	1	113
0	0	2000
1	1	94
1	0	2000

As we can see from the output of the SAS program, we see that physical activity is negatively confounding the effect of gender (the log ratio is larger when analyzing the stratified samples). Thus, if we see the effect of gender in a bivariate approach, we will not see the effect of gender. This is further confirmed by a logistic regression model.

```
data data1;
   input phy_act gender resp count;
   cards;
   0 1 1 84
   0 1 0 1200
   0 0 1 80
   0 0 0 800
   1 1 1 10
   1 1 0 800
   1 0 1 33
   1 0 0 1200
```

```
    ;
run;
data data1(keep=phy_act gender resp);
   set data1;
   do i=1 to count;
      output;
   end;
run;

proc freq data=data1;
   table resp*gender/chisq;
run;


proc freq data=data1;
   table resp*gender/chisq;
   by phy_act;
run;

/* Multivariate Analysis */
proc logistic data=data1 descending;
   model resp=gender phy_act;
run;
```

Statistics for Table of resp by gender

Statistic	DF	Value	Prob
Chi-Square	1	1.6582	0.1978
Likelihood Ratio Chi-Square	1	1.6606	0.1975
Continuity Adj. Chi-Square	1	1.4797	0.2238
Mantel-Haenszel Chi-Square	1	1.6578	0.1979
Phi Coefficient		-0.0199	
Contingency Coefficient		0.0198	
Cramer's V		-0.0199	

phy_act=0

Statistics for Table of resp by gender

Statistic	DF	Value	Prob
Chi-Square	1	4.8431	0.0278
Likelihood Ratio Chi-Square	1	4.7720	0.0289
Continuity Adj. Chi-Square	1	4.4860	0.0342
Mantel-Haenszel Chi-Square	1	4.8408	0.0278
Phi Coefficient		-0.0473	
Contingency Coefficient		0.0473	
Cramer's V		-0.0473	

phy_act=1

Statistics for Table of resp by gender

Statistic	DF	Value	Prob
Chi-Square	1	4.9323	0.0264
Likelihood Ratio Chi-Square	1	5.2925	0.0214
Continuity Adj. Chi-Square	1	4.2573	0.0391
Mantel-Haenszel Chi-Square	1	4.9299	0.0264
Phi Coefficient		-0.0491	
Contingency Coefficient		0.0491	
Cramer's V		-0.0491	

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq

Intercept	1	-2.2631	0.1097	425.3395	<.0001
gender	1	-0.4339	0.1464	8.7801	0.0030
phy_act	1	-1.4254	0.1768	65.0002	<.0001

In this case, the response rate is higher for female among two subpupulations defined by physical active indicator. When we pool the data together, due to the negative confounding effect of physical activity on gender, the effect of gender is reduced. However, the overall response rate (without controling physical activity) is still higher for females. In extreme cases, we may encounter the situation where the effect of a factor is reversed when we combine the population. This is known as Simpson' paradox.

Consider the following example

Urban/Rural	Urban	Urban	Rural	Rural
	No effect	Cure	No effect	Cure
Standard drug	500	100	350	350
New Druge	1050	350	120	180

It is easy to verify that the new drug does better in both urban and rural area. Yet, for the pooled data, the standard drug gives a higher cure rate!

It should be clear at this point why bivariate analysis can lead to biased results. Unfortunately, this is the fact we have to live with. This is especially true in observational study. Imagine we never observe these confounders in the database, our model is necessarily biased or misspecified. In this case, we would mistakenly filter out an otherwise important variable and therefor make a type II error.

Another type of error is mistakenly assume the importance of a particular predictor while in reality, its "predictiveness" is solely based on its correlation with other variables.

Let's consider the following SAS program (This is taken from Bertrand (1998)):

```
data data1;
input x1 x2 x3 x4 x5 y;
cards;
```

87 84 217 133 37 76
236 155 180 290 177 243
91 111 83 149 52 20
164 10 145 244 281 139
121 167 120 72 73 42
17 87 31 114 99 10
137 189 112 167 97 130
82 118 82 10 252 37
85 41 150 218 167 107
166 136 241 254 61 237
208 218 156 131 182 205
165 107 240 225 167 202
102 41 147 234 217 143
83 228 58 151 10 96
267 104 190 280 190 209
160 69 203 135 100 67
167 137 108 128 198 86
109 135 61 100 270 88
248 226 94 169 169 172
10 107 75 135 13 41
237 246 15 104 290 155
222 249 108 202 266 284
89 128 83 192 49 88
129 145 83 67 215 56
145 37 205 147 165 131
210 214 220 124 103 169
175 132 165 266 41 154
137 236 58 37 27 64
219 284 116 49 261 186
276 156 290 209 238 276
176 220 171 86 160 193
260 210 114 209 133 215
101 36 193 176 137 75
219 165 225 181 178 222
217 251 85 176 115 193
283 191 220 170 237 241
35 187 86 58 45 40
176 272 73 46 131 123

```
68 194   116   65 23 60
78 180   56 62 47 22
100   103   80 113   180   58
290   197   252   274   148   290
157   129   172   200   53 148
139   206   68 167   153   135
155   249   111   70 177   174
128   101   172   203   65 96
153   197   183   190   20 154
173   201   10 141   273   166
134   156   140   91 95 97
240   290   99 223   68 217
;

proc reg data=data1;
   model y=x1 x2 x3 x4 x5;
run;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-193.93876	14.89732	-13.02	<.0001
x1	1	-0.08741	0.08378	-1.04	0.3025
x2	1	0.77530	0.06725	11.53	<.0001
x3	1	0.53391	0.05584	9.56	<.0001
x4	1	0.62394	0.05197	12.01	<.0001
x5	1	0.38579	0.04270	9.04	<.0001

It can be easily verified that x_1 has the highest bivariate correlation with y and yet it is insignificant in the multivariate analysis.

It can further be shown that including x_1 and removing any other variables will reduce the R^2 significantly. Thus, importance of x_1 is mainly due to its correlation with other predictors. If we calculate a partial correlation, we find that given x_2, x_3, x_4, x_5, the correlation between y and x_1 is no longer

significant. Indeed, the indepedent contribution of each variable is based on the partial correlation. (proc proc corr data=data1; partial x2 x3 x4 x5; var y;with x1; run;).

Do a factor analysis on this problem with y and without y!

Having analyzing the problems with bivariate screening, with due caution, it is still one of most important and pratical approaches for initial variable selection. We shall devote several subsections on this topic.

5.2 Generalized Correlation

When both y and x are continuous variables, the most commonly used association measure is still Pearson's correlation $r_{x,y}$. It is important to point out that Pearson's correlation measures the linear relationship between x and y. Spearman's rank correlation is more appropriate when there exists a general monotone relationship between x and y which is not necessary linear. In any data, we would expect both continuous and discrete variables. Discrete variables can be further classified as ordinal and nominal. For ordinal variable, the regular correlations still apply. We present some generalized correlations that deal with nominal variables. These definitions have been widely used in data mining and machine learning community.

- **y is continous and x is nominal:** Suppose x has a probability distribution $\{x^i, p_x^i\}, i = 1, 2, \ldots k$. (Here, we do not make the further differentiation between nomial and ordinal. As we shall see, we are essentially fitting a saturated model and such differentiation is not necessary.) Notice, here we explicitly assume that $k \ll n$ where n is the cardinality of x. We define $x_{bi} = 1$ whenever x takes the value x^i, 0 otherwise. Then, the generalized correlation is defined as

$$r_{x,y} = \sum_{i=1}^{k} p_x^i |r_{y,x_{bi}}|. \tag{13}$$

  The following SAS macro implements the algorithm. As you can see from the example, it is obvious that x should be more predictive than w and this is reflected by the calculated generalized correlation coefficients.

```
%macro hwgencorr1(indat,y,x);
```

```
data _hwgen_temp1;
   set &indat(keep=&y &x);
run;
data _null_;
   if 0 then set _hwgen_temp1 nobs=totobs;
   call symput ("totobs", left(put(totobs,8.0)));
   stop;
run;

proc sql;
   create table _hwgen_temp2 as
   select &x,
          count(*)/&totobs as px
   from _hwgen_temp1
   group by &x;
quit;

proc sql noprint;
   select &x into : xvalues separated by ' '
   from _hwgen_temp2;
quit;

data _null_;
   if 0 then set _hwgen_temp2 nobs=x_cnt;
   call symput ("x_cnt", left(put(x_cnt,8.0)));
   stop;
run;

data _hwgen_temp2;
   set _hwgen_temp2;
   _name_=compress("xb"||_n_);
run;

data _hwgen_temp1;
   set _hwgen_temp1;
   %do I=1 %to &x_cnt;
      %let xvalue=%scan(&xvalues,&I, %str( ));
      xb&i=(&x="&xvalue");
```

```
   %end;
run;

proc corr data=_hwgen_temp1
          outp=_hwgen_temp3(where=(_type_='CORR')) noprint;
   var &y;
   with xb:;
run;

proc sql;
   select sum(abs((a.&y))*(b.px)) as gencorr1
   from _hwgen_temp3 as a,
        _hwgen_temp2 as b
   where a._name_=b._name_;
quit;

%mend;

data check;
   do i=1 to 100;
      y=ranuni(10);
      if y<=0.2 then x="A";
      else if y<=0.6 then x="B";
      else x="C";
      t=ranuni(20);
      if t<0.3 or t>0.9 then w='A';
      else if t>0.3 and t<0.7 then w='B';
      else w='C';
      output;
   end;
run;

%hwgencorr1(check,y,w);
%hwgencorr1(check,y,x);
```

- **Both y and x are nominal:** Suppose x, y has probability distributions $\{x^i, p_x^i\}, i = 1, 2, \ldots k$ and $\{y^j, p_y^j\}, j = 1, 2, \ldots l$ respectively.

Again, we define x_{bi} as an indicator when x takes the value x^i and y_{bj} as an indicator when y takes the value y^j. The generalized correlation is defined as

$$r_{x,y} = \sum_{i=1}^{k} \sum_{j=1}^{l} p(x = x^i, y = y^j) |r_{y_{bj}, x_{bi}}|. \tag{14}$$

Notice in this case, the association can also be measured by Chi-Square. For some comparison on these two different measures, see ???

5.3 Weight of Evidence

Consider data matrix $(Y|X)$ of $n \times 2$,where Y is a binary variable and X is a categorical variable with distinct values $X^1, X^2, \ldots X^K$. A common statistics to assess the association of X and Y is Chi-Square. In this section, we introduce another common measure of association. Furthermore, as we will see from the construction, this statistics can be used to measure the discrimatory power of the categorical variable X.

Given a value X^i, let p^i be the proportion of observations where $X = X^i$ and $Y = 1$ and q^i be the proportion of observations where $X = X^i$ and $Y = 0$. The weight of evidence is defined as $ln(p^i/q^i)$. The association statistics is defined as $\sum_i^K (p^i - q^i) ln(p^i/q^i)$. We illustrate the concept using an example. Considering the following standard cross tabulation.

	$y = 1$	$y = 1$
$x = 1$	n_1	n_2
$x = 2$	n_3	n_4
$x = 3$	n_5	n_6

Here, $p^1 = \frac{n_1}{n_1+n_2}$ and $q^1 = \frac{n_2}{n_1+n_2}$. Thus $w^1 = In(\frac{p^1}{q^1}) = In(\frac{n_1}{n_2})$.
The association statistics is defined as

$$\sum_{1}^{3} \frac{n_i - n_{i+1}}{n_i + n_{i+1}} In(\frac{n_i}{n_{i+1}}). \tag{15}$$

This statistics is only applicable when the Y variable is binary. In the case when Y is multinomial, the calculation is more complicated.

5.4 Information Theory Based Condition

We briefly introduce Shannon's measure of average information contained in an experiment for which the outcome may be one of the c categories. The amount of information associated with an event which has probability p can be measured by the quantity $log(1/p)$. This measure should be intuitive in the sense that the more unlikely the event, the greater the surprise when that event actually does occur, and so the more information is provided by the knowledge that the event has occurred. The presence of the logarithm ensures that the information is additive. (Logarithm has lots of nice features. For example, $log(a/b)$ is close to $(a^2 - b^2)/2ab$ when a/b is close to 1. This ensures the likelihood ratio test is approximatedly Chi-Square.)

Suppose an experiment can have one of c mutually exclusive, exhaustive outcomes, which have respective probabilities $P(i), i = 1, 2, \ldots, c$, then the average amount of information provided by that experiement is given by

$$H = E(log(p)) = -\sum_{i=1}^{c} P(i) Log(P(i)). \tag{16}$$

H is called Shannon measure of information. It is also called entropy of average uncertainty. The maximum possible value of H is equal to $log(c)$ and this occurs when the outcomes are equally probable (discrete uniform).

Let y denote a class variable with values $y^1, y^2, \ldots, y^k$ and x is a discrete predictor with values $x^1, x^2, \ldots, x^l$(order is not assumed or utilized). Let's consider the conditional entropy

$$H(y|x) = -\sum_{i=1}^{l} p(x = x^i)[\sum_{j=1}^{k} p(y = y^j|x = x^i) log(p(y = y^j|x = x^i))]. \tag{17}$$

The conditional entropy has a very distinct meaning, namely, the entropy (randomness) of y that is not removed by knowing the values of x. The smaller $H(y|x)$, the better we can say about y using the information of x. If $H(y|x) = 0$, that implies value of y is fixed when we know x.

$M(x, y) = H(y) - H(y|x) = H(x) - H(x|y) = H(x) + H(y) - H(x, y)$ is defined as the information gain by knowing x. This measure biases towards x with more distinct values. Thus, we define the following normalized measure called uncertainly ratio

$$UR(x,y) = 2 \times \frac{H(y) - H(y|x)}{H(y) + H(x)}. \tag{18}$$

Here are SAS macros to calculate these measures.

```
%macro entropy(indat,catvar,totobs,val);
%global &val;
proc sql;
   select -sum(prob*log(prob)) into: &val
   from (    select count(*)/&totobs as prob,
                     &catvar
             from &indat
             group &catvar);
quit;
%mend;


/* Conditional Entropy H(Y|X) */

%macro condentropy(indat,target,catvar,totobs,val);
%global &val;

proc sql;
   create table _entropy_t as
      select count(*) as count,
             &target,
             &catvar
      from &indat
      group &catvar, &target
      order by &catvar;

   create table _entropy_t as
      select &catvar,
             &target,
             count,
             sum(count) as totperx,
             count/calculated totperx as proby_x
```

```
      from _entropy_t
      group by &catvar;

      select sum( -proby_x*log(proby_x)*(totperx/&totobs)) into: &val
      from _entropy_t;
quit;

%mend;

data check;
   input x y count;
   cards;
   1 1 10
   1 2 20
   2 1 30
   2 2 40
   ;
run;

data check (keep=x y);
    set check;
    do i=1 to count;
       output;
    end;
run;

%entropy(check,y,100,ans1);
%entropy(check,x,100,ans2);
%condentropy(check,y,x,100,ans3);
data _null_;
   ur=2*(ans1-ans3)/(ans1+ans2);
run;
```

In this case, the symmetric uncertainly coefficient is .006265403, which is very small. Indeed, if we do a Chi Square association test on x, y, we see there is little association.

In the case of continous variable, one needs to replace summation with

integral. It is not practical to calculate entropy of continuous variales using sample data since we do not know the distributions in advance. However, it is obvious that we can use entropy to assess assocation between two categorical variables. In a latter section, we will also show how this idea can be used to discretize continuous variables.

5.5 Zeta Association

This is a statistics that can be used to measure the association between two binary variables.

Consider the following cross tabulation of y and x:

y/x	$y=0$	$y=1$
$x=0$	n_1	n_2
$x=1$	n_3	n_4

Here, we use the value of x to predict the value of y. If we think of prediction in terms of rules (This is used in decision tree!), there are only two rules $[x=0 \rightarrow y=0, x=1 \rightarrow y=1]$ or $[x=0 \rightarrow y=1, x=1 \rightarrow y=0]$. For a given rule, we can calculate its accuracy based on the data presented in the contigency table. For example, if the first rule is used, we would get n_1+n_4 out of $n_1+n_2+n_3+n_4$ cases classified correctly. Zeta is defined as the precentage of classification accuracy if the best rule is used for prediction. In a sense, this number gives the maximum potential of a variable in using for prediction. Mathematically, it is defined as

$$Z = \frac{\max(n_1+n_4, n_2+n_3)}{n_1+n_2+n_3+n_4}. \tag{19}$$

It is possible to generalize the definition to high level contigency tables. A good paper on this subject is *Zeta: A Global Method for Discritization* by Ho and Scott.

This statistics is a member of so called prediction efficacy index. The advantage of them over traditional Chi Square association is that they are strictly prediction oriented. Thus, a higher Chi Square value does not necessarily imply a higher classification accuracy rate. A higher Z value would imply that this variable will improve the classification accuracy rate better than another variable with a lower Z value. (Obviously, it all depends on

the optimization algorithms used to derive the rules. Lots of regression techniques are designed to minimize error, which in same cases is not equivalent to classification accuracy rate.)

Let's consider an example. Given the following data

y/x	$x_1 = 0$	$x_1 = 1$		$x_2 = 0$	$x_2 = 1$
$y = 0$	10	300		260	50
$y = 1$	100	20		40	80

It is trivial to verify that $Z(x_1|y) > Z(x_2|y)$. We therefore expect x_1 a better predictor than x_2. We perform a logistic regression on each predictor individually and use $p = 0.5$ as the cutoff point for classification.

```
data data1;
  input y x count xind;
  cards;
  0 0 10  1
  0 1 300 1
  1 0 100 1
  1 1 20  1
  0 0 260 2
  0 1 50 2
  1 0 40 2
  1 1 80 2
  ;
run;
proc logistic data=data1 descending;
   freq count;
   model y=x /ctable pprob=0.5;
   by xind;
run;
```

The selected output suggests that x_1 does a better job in classification than x_2.

xind=1

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.3026	0.3317	48.1991	<.0001
x	1	-5.0106	0.4041	153.7123	<.0001

Classification Table

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.500	100	300	10	20	93.0	83.3	96.8	9.1	6.3

xind=2

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.8717	0.1698	121.4571	<.0001
x	1	2.3417	0.2477	89.3916	<.0001

Classification Table

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.500	80	260	50	40	79.1	66.7	83.9	38.5	13.3

As we will see in a subsequent section, a more fruitful use of Z measure is to discretize a continuous variable in a way such that Z is maximized.

5.6 Assessing Association between two Ordinal Variables

Chi-Square test for independenc is sometimes employed to detec the association. In this case, if we observe large Chi-Square statistics with small p value, we can conclude that there are associations that are statistically significant. However, using Chi-Square for ordinal varaibles is not efficient in the sense that the test does not make use of the ordinal information about the data. This point is clear because the chi-square test is permutationally invariant.

5.7 Other Traditional Association Test

In this section, we introduce some standard traditional association test for ordinal varibles. They include Kendall's tau-b (τ_b), Stuart's tau-c (τ_c), Somers' D_{yx} and Goodman and Kruskal's gamma γ.

First, define a two-way $R \times C$ table that represents the cross-classification of two ordinal variables X and Y. Let X be the row variable with R categories and Y be the column variable that has C ordered categories. Furthermore, let p_{ij} be the cell probability of the ith row and jth column, p_{i+} and p_{+j} be the marginal probability of the ith row and jth column, respectively, $i = 1, \ldots, R, j = 1, \ldots, C$. Define the following terms:

$$A_{ij} = \sum_{k>i}^{R} \sum_{l>j}^{C} p_{kl} + \sum_{k<i}^{R} \sum_{l<j}^{C} p_{kl}, D_{ij} = \sum_{k>i}^{R} \sum_{l<j}^{C} p_{kl} + \sum_{k<i}^{R} \sum_{l>j}^{C} p_{kl}, \qquad (20)$$

$$\Pi_a = \sum_{i=1}^{R}\sum_{j=1}^{C} p_{ij}D_{ij}, \Pi_d = \sum_{i=1}^{R}\sum_{j=1}^{C} p_{ij}D_{ij}. \tag{21}$$

$$H = \sum_{i=1}^{R}\sum_{j=1}^{C} p_{ij}(A_{ij} - C_{ij})^2. \tag{22}$$

Here Π_a is the proabability that a pair of observations has concordant orders, and Π_d is the probability that a pair of observations has discordant orders.

Goodman and Krustal's *gamma*(γ) is defined as

$$\gamma = \frac{\Pi_a - \Pi_d}{\Pi_a + \Pi_d}. \tag{23}$$

. It is a measure of the difference between the c onditional probabilities of concordant and discordant pairs of observations given that they are not tied on either X or Y.

Kendall's tau-b (τ_b) is similar to gamma except that it has a correction for ties. The formula is

$$\tau_b = \frac{\Pi_a - \Pi_d}{\sqrt{(1 - \sum_{i=1}^{R} p_{i+}^2)(1 - \sum_{j=1}^{C} p_{+j}^2)}}. \tag{24}$$

Stuart's tau-c (τ_c) makes an adjustment for table size and is defined as

$$\tau_c = \frac{\Pi_a - \Pi_d}{1 - m^{-1}}, \tag{25}$$

where $m = min(R, C)$.

Somers's D_{yx} is an asymmetric measure of association in which the column variable Y is regarded as the dependent variable and the row variable X is regarded as independent variable. D_{yx} is similar to τ_b, except that D_{yx} makes a correction for ties only on the independent variable. It is defined as

$$D_{yx} = \frac{\Pi_a - \Pi_d}{1 - \sum_{i=1}^{R} p_{i+}^2}. \tag{26}$$

5.8 Several SAS Macros for Variable Screening

We introduce several SAS macros that are designed to initially screen the preliminary predictive power of each variable. Notice we have mentioned that technically, such approach is not appropriate and does not lead to optimal models. However, when we face with thousands of variables, this kinds of screening routines can be very useful as the first step of a data reduction process.

First, we discuss the case when the outcome variable y is continuous. In this case, we assume that a linear regression model will be used to fit the data. Among the predictors, we assume there exist both continuous and categorical variables. If x is categorical, we fit a saturated model by using x as a class variable. If x is continuous, we fit a polynomial $y = f(x) + \epsilon$ where $f(x)$ is a polynomial function upto order 3. The idea is that we use this routine as a pre-processing step. As a result, we will not be able to manually access the suitable transformation and functional form for each predictor. However, since any function can be reasonably approximated by a polynomial of order 3 (due to Taylor Theory), we believe that our approach is flexible enough. It is important to point out that we are not concerned with the issue of overfitting at this point. We simply screen each variable in a semi-automated way by asking the following question : "Does this variable have any potential to be useful at all"? If the answer is no, this variable will not be included in the subsequent variables. If the answer is yes, it is up to the user at the next step to come up with more refined transformation to deal with possible missing values, nonlinearity and other issues.

For each variable, the global F statistics and its associated P value are collected. Also, the R^2 statisitcs is also included. Once this routine is performed, a dataset containing all the variables and their corrresponding measures of association and predictiveness will generated. A user can then query this dataset to select variables that will be kept to the next stage of the analysis.

```
*************************************************************************;
* Program: hwrsqsrch.sas;
* Author: Hongjie Wang;
* Purpose: Search important predictors by R^2 , Global F;
* Syntax: %hwrsqrch(indat,y,outdat,order);
```

```
* indat: input dataset;
* y: the outcome variable, continuous, we assume linear regression;
* We assume that other than y, every other numeric var is a predictor;
* outdat: the results of the analysis;
* It has the following layout [var_name, rsquare, Fvalue, ProbF];
* For each variable x, we fit a model y=x x^2 x^3;
* If x is categorical var (char), we fit a saturated model - ANOVA;
* The resulting Rsquare, Global F stat and its associated P value;
* order: if x continuous, we allow the control of ;
* order of polynomial y=f(x);
* The idea is that we do not know the exact functional form;
* and a third order polynomial might be flexible enough to ;
* approximate the functional form;
* can be rank ordered to access the potential of each variable;
* This macro can be used as a pre-processing, screening step;
* Date: Jan 16, 2002 to celebrate my own birthday;
* Y=f(x1,x2,x3...);
* y is continuous and x1, x2 .. continuous or categorical;
* It is assumed that if x is numerical, then it is continuous;
* and if x is char, then it is categorical;
*********************************************************************;



%macro hwrsqsrch(indat,y,outdat,order);
ods trace off;
ods select none;
 data &outdat;
   length var_name $40;
   var_name=' ';
   Rsqrt=.;
   Fvalue=.;
   probf=.;
run;

proc contents data=&indat(drop=&y)
   out=varnames(keep=type name) noprint;
```

```
run;

proc sql noprint;
   select name into : char_varlist separated by ' '
   from varnames
   where type=2;
   select count(distinct name) into :char_varnbr separated by ' '
   from varnames
   where type=2;
   select name into : num_varlist separated by ' '
   from varnames
   where type=1;
   select count(distinct name) into : num_varnbr separated by ' '
   from varnames
   where type=1;
quit;


%do i=1 %to &num_varnbr;
   %let var_cur=%scan(&num_varlist,&i);
   ods output overallAnova=hwrsqsrch_temp1;
   ods output fitstatistics=hwrsqsrch_temp2;
   %let exp=&var_cur;
   %if &order=2 %then %do;
      %let exp=&var_cur &var_cur*&var_cur;
   %end;
   %if &order=3 %then %do;
      %let exp=&var_cur &var_cur*&var_cur &var_cur*&var_cur*&var_cur;
   %end;
   proc glm data=&indat(keep=&y &var_cur);
      model y=&exp;
   run;
   data _null_;
      set hwrsqsrch_temp2;
      call symput ("rsquare", left(put(RSquare,8.4)));
   run;

   data _null_;
```

```
      set hwrsqsrch_temp1;
         if Source='Model' then do;
            call symput("Fvalue", left(put(fvalue,8.3)));
            call symput("ProbF",  left(put(probf,8.4)));
         end;
   run;

   data hwrsqsrch_temp3;
      length var_name $40;
      var_name="&var_cur";
      Rsqrt=&rsquare;
      fvalue=&Fvalue;
      probf=&ProbF;
   run;

   data &outdat;
      set &outdat hwrsqsrch_temp3;
   run;
%end;

%do i=1 %to &char_varnbr;
   ods output overallAnova=hwrsqsrch_temp1;
   ods output fitstatistics=hwrsqsrch_temp2;

   %let var_cur=%scan(&char_varlist,&i);
   proc glm data=&indat(keep=&y &var_cur);
      class &var_cur;
      model &y=&var_cur;
   run;
   data _null_;
      set hwrsqsrch_temp2;
      call symput ("rsquare", left(put(RSquare,8.4)));
   run;

   data _null_;
      set hwrsqsrch_temp1;
         if Source='Model' then do;
            call symput("Fvalue", left(put(fvalue,8.3)));
```

```
            call symput("ProbF",  left(put(probf,8.4)));
         end;
   run;

   data hwrsqsrch_temp3;
      length var_name $40;
      var_name="&var_cur";
      Rsqrt=&rsquare;
      fvalue=&Fvalue;
      probf=&ProbF;
   run;

   data &outdat;
      set &outdat hwrsqsrch_temp3;
   run;
%end;


data &outdat;
   set &outdat;
   if compress(var_name)^='';
run;

ods select all;
%mend;

/*
*Test Exampe;

data check;
    do i=1 to 100;
       x1=ranuni(10);
       x2=ranuni(100);
       y=x1*1.3+2.5*x1*x1+4.5*x2+rannor(1000);
       if y<0 then z='a';
       else if y<2 then z='b';
       else z='c';
       if ranuni(1000)<0.45 then w='bb';
```

```
        else w='aa';
        output;
    end;
run;



options mprint symbolgen;

%hwrsqsrch(check,y,results,1);
proc print data=results;
run;

%hwrsqsrch(check,y,results,2);
proc print data=results;
run;

%hwrsqsrch(check,y,results,3);
proc print data=results;
run;

*/
```

When y is categorical, we have three different situations: y is binary, ordered or multi-nomial. When y is binary, an ordinary logistic regression model is used. If y is ordered, we fit the data with a cumulative logit model. When y is nomial with more than 2 categories, we fit a generalized logit model. Again, if x is numeric and continuous, we optinally allow higher order terms in the model. The global log likelihodd ratio χ^2 and its p value is calcuated for each predictor.

```
*************************************************************************;
* Program: hwlogitsrch.sas;
* Author: Hongjie Wang;
* Purpose: variable selection, screening routine for categorical outcome;
* %hwlogitsrch(indat,classvar,outdat,type,order);
```

```
* indat: input dataset;
* classvar: the categorical outcome variable,numerical;
* It can be binary, ordinal or nominal;
* outdat: the results are summarized in this dataset;
* It contains the following:;
* var_name: predictor;
* lrchi: the likelihood ratio chisq statistics;
* lrchip: the p value assocated with this statistics;
* type: describe the type of distribution of the classvar;
* It takes the following types: bin, ord, mult;
* order: if the predictor is numerical and continuous;
* the user can control the order of polynomial to fit;
* for binary, a logistic regression is fit;
* for ordinal outcome, a cumulative logit model is fit;
* for multinomial outcome, a generalized logit model is fit;
* if the predictor is categorical (char), it is treated as class var;
* the likelihood ratio chisq is the difference between;
* fit model and the null model, thus,the larger the more predictive;
*********************************************************************;



%macro hwlogitsrch(indat,classvar,outdat,type,order);

ods trace off;
ods select none;
 data &outdat;
   length var_name $40;
   var_name=' ';
   LRchi=.;
   LRchiP=.;
run;

proc contents data=&indat(drop=&classvar)
   out=varnames(keep=type name) noprint;
run;

proc sql noprint;
```

```
   select name into : char_varlist separated by ' '
   from varnames
   where type=2;
   select count(distinct name) into :char_varnbr separated by ' '
   from varnames
   where type=2;
   select name into : num_varlist separated by ' '
   from varnames
   where type=1;
   select count(distinct name) into : num_varnbr separated by ' '
   from varnames
   where type=1;
quit;


%do i=1 %to &num_varnbr;
   %let var_cur=%scan(&num_varlist,&i);
   ods output GlobalTests=hwlogitsrch_temp1;
   %let exp=&var_cur;
   %if &order=2 %then %do;
      %let exp=&var_cur &var_cur*&var_cur;
   %end;
   %else %if &order=3 %then %do;
      %let exp=&var_cur &var_cur*&var_cur &var_cur*&var_cur*&var_cur;
   %end;
   proc logistic data=&indat(keep=&classvar &var_cur);
      model &classvar=&exp %if &type=mult %then
                                    /link=glogit;;
   run;
   data _null_;
      set hwlogitsrch_temp1;
      if test='Likelihood Ratio' then do;
         call symput ("lrchi", left(put(chisq ,8.4)));
         call symput ("lrchip", left(put(probchisq,8.4)));
      end;
   run;
```

```
   data hwlogitsrch_temp2;
      length var_name $40;
      var_name="&var_cur";
      lrchi=&lrchi;
      lrchip=&lrchip;
   run;

   data &outdat;
      set &outdat hwlogitsrch_temp2;
   run;
%end;

%do i=1 %to &char_varnbr;
   ods output GlobalTests=hwlogitsrch_temp1;
   %let var_cur=%scan(&char_varlist,&i);
   proc logistic data=&indat(keep=&classvar &var_cur);
      class &var_cur;
      model &classvar=&var_cur %if &type=mult %then
                                   /link=glogit;;
   run;
   data _null_;
      set hwlogitsrch_temp1;
      if test='Likelihood Ratio' then do;
         call symput ("lrchi", left(put(chisq ,8.4)));
         call symput ("lrchip", left(put(probchisq,8.4)));
      end;
   run;


   data hwlogitsrch_temp2;
      length var_name $40;
      var_name="&var_cur";
      lrchi=&lrchi;
      lrchip=&lrchip;
   run;

   data &outdat;
      set &outdat hwlogitsrch_temp2;
```

```
   run;


%end;


data &outdat;
   set &outdat;
   if compress(var_name)^='';
run;

ods select all;
%mend;

* test example;

/*
data check;
    do i=1 to 100;
       x1=ranuni(10);
       x2=ranuni(100);
       y=x1*1.3+2.5*x1*x1+4.5*x2+rannor(1000);
       if y<0 then multy='a';
       else if y<1 then multy='b';
       else multy='c';
       if y<0 then ordy=1;
       else if y<1.5 then ordy=2;
       else ordy=3;
       if x2<0.3 then c='level1';
       else if x2<0.7 then c='level2';
       else c='level3';
       biny=(y<2);
       output;
    end;
run;

options mprint;
%hwlogitsrch(check,multy,result,mult,1);
```

```
proc print data=result;
run;
%hwlogitsrch(check,multy,result,mult,2);
proc print data=result;
run;
%hwlogitsrch(check,multy,result,mult,3);
proc print data=result;
run;


%hwchisrch(check,multy,result,2);
proc print data=result;
run;
%hwchisrch(check,multy,result,3);
proc print data=result;
run;

%hwchisrch(check,multy,result,4);
proc print data=result;
run;


%hwlogitsrch(check,biny,result,bin,1);
proc print data=result;
run;
%hwlogitsrch(check,biny,result,bin,2);
proc print data=result;
run;
%hwlogitsrch(check,biny,result,bin,3);
proc print data=result;
run;


%hwchisrch(check,biny,result,2);
proc print data=result;
run;
%hwchisrch(check,biny,result,3);
proc print data=result;
```

```
run;

%hwchisrch(check,biny,result,4);
proc print data=result;
run;


%hwlogitsrch(check,ordy,result,ord,1);
proc print data=result;
run;
%hwlogitsrch(check,ordy,result,ord,2);
proc print data=result;
run;
%hwlogitsrch(check,ordy,result,ord,3);
proc print data=result;
run;


%hwchisrch(check,ordy,result,2);
proc print data=result;
run;
%hwchisrch(check,ordy,result,3);
proc print data=result;
run;

%hwchisrch(check,ordy,result,4);
proc print data=result;
run;

*/
```

When y is categorical, χ^2 can be used to measure the association between y and x. If x is continuous, one can first discretize it into groups. The following SAS macro streamlines the process of calculating χ^2 of each predictor and y.

```
*****************************************************************;
* Program: hwchisrch.sas;
* Author: Hongjie Wang, 1/30/02;
* Purpose: Variable selection by chisq;
* %hwchisrch(indat,classvar,outdat,group);
* indat: input dataset;
* classvar: the categorical outcome var;
* outdat: resulting dataset contains the following information;
* var_name: variable name;
* chisq: chisq of the predictor and the outcome var;
* chisq_pvalue: p value assoicated with chisq;
* High Chisq corresponds to high degree of association;
* When the predictor is numeric, it is discretized into groups;
* group: number of groups of the resulting discretized var;
*****************************************************************;

 %macro hwchisrch(indat,classvar,outdat,group);

data &outdat;
   length var_name $40;
   var_name="";
   chisq=.;
   chisq_pvalue=.;
run;


proc contents data=&indat(drop=&classvar)
   out=varnames(keep=type name) noprint;
run;

proc sql noprint;
   select name into : char_varlist separated by ' '
   from varnames
   where type=2;
   select count(distinct name) into :char_varnbr separated by ' '
   from varnames
   where type=2;
```

```
   select name into : num_varlist separated by ' '
   from varnames
   where type=1;
   select count(distinct name) into : num_varnbr separated by ' '
   from varnames
   where type=1;
quit;

%do i=1 %to &char_varnbr;
   %let var_cur=%scan(&char_varlist,&i);
   proc freq data=&indat(keep=&classvar &var_cur) noprint;
      table &classvar*&var_cur /chisq;
      output out=hwchisrch_temp1(keep=_PCHI_ P_PCHI ) chisq;
   run;

   data hwchisrch_temp1;
      set hwchisrch_temp1;
      length var_name $40;
      var_name="&var_cur";
      rename _pchi_=chisq;
      rename p_pchi=chisq_pvalue;
   run;
   data &outdat;
      set &outdat hwchisrch_temp1;
   run;
%end;

%do i=1 %to &num_varnbr;
   %let var_cur=%scan(&num_varlist,&i);
   proc rank data=&indat(keep=&var_cur &classvar) group=&group
      out=hwchisrch_temp1;
      var &var_cur;
      ranks hwchisrch_rankx;
   run;


   proc freq data=hwchisrch_temp1 noprint;
      table &classvar*hwchisrch_rankx /chisq;
```

```
      output out=hwchisrch_temp1(keep=_PCHI_ P_PCHI ) chisq;
   run;

   data hwchisrch_temp1;
      set hwchisrch_temp1;
      length var_name $40;
      var_name="&var_cur";
      rename _pchi_=chisq;
      rename p_pchi=chisq_pvalue;
   run;
   data &outdat;
      set &outdat hwchisrch_temp1;
   run;
%end;

data &outdat;
   set &outdat;
   if compress(var_name)^='';
run;

%mend;
/*
* test case;

data check;
   do i=1 to 100;
      x1=ranuni(10);
      x2=ranuni(20);
      if x1<0.23 or x1>0.9 then y=1;
      else y=0;
      r=ranuni(45);
      if r<0.3 then z1='a';
      else if r<=.8 then z1='b';
      else z1='';
      if r<0.5 then z2='cc';
      else z2='dd';
      output;
   end;
```

```
run;

options mprint symbolgen;

%hwchisrch(check,y,result,2);
proc print data=result;
run;

%hwchisrch(check,y,result,3);
proc print data=result;
run;
%hwchisrch(check,y,result,4);
proc print data=result;
run;
*/
```

6 Categorization and Discretization

Categorization refers to the process of collapse a categorical variable with K distinct values into a categorical variable with M distinct values, where $M \ll K$. There might be additional constraints on group to preserve the order structure if the original categorical variable is an ordered one. Discretization converts a continuous variable into a discrete variable (possibly ordered) with few categories. They are two special cases of homogenous grouping operation.

Homogenous grouping occur in lots of applications, see Wang (2001) for a review. In particular, most of the data mining and machine learning algorithms have pre-processing steps to convert all variables into categorical variables with manageable number of values. It is intuitive that, by performing such operation, some information contained in the original variables is necessarily lost. The optimization algorithms behind most of the grouping techniques try to minimize such loss of information. There are benefits for paying such price on information loss. In many tree models, discretization significantly reduces the computational complexity. In statistical models, by reducing the number of categories, degree of freedom is usually reduced accordingly. In addition, grouping is an viable alternative to transformation in dealing with nonlinearity.

There are two classes of grouping procedures, supervised and unsupervised. The unsupervised case, binning is the most commonly used discretization technique. We briefly outline an unsupervised grouping technique based on distribution and then devote most of the section on various supervised grouping techniques.

6.1 Optimum Grouping for a Continuous Variable

Given continously distributed random variable x with distribution function $f(x)$, mean μ and variance σ^2, we want to divide the values into K groups. The objective is to miminize the loss of information. There are two sets of parameters we need to decide, the number of groups K and the boundary cutoff points. One of the measure of information loss is given the

$$\text{correlation}(x, d(x)),$$

where $d(x)$ is the discretized version of the variable.

This translates to

$$\sum_{i=1}^{K} \int_{a_{i-1}}^{a_i} (x - \mu_i)^2 f(x) dx. \tag{27}$$

Here, μ_i is the conditional mean of the interval $[a_{i-1}, a_i]$.

The above is a standard optimization problem. It can be shown that $a_i = (\mu_i + \mu_{i+1})/2$. Thus, the boundary between groups i and $i+1$ is the mean of the two means of these groups.

Suppose we arrange the observations in ascending rank order and a frequency distribution is formed. Let n be the total number of observations and n_i represent the number of observations in group i. Then, given K, it can be shown the we are essentially trying to minimize the following:

$$S = \sum_{i=1}^{K} n_i \sigma_i^2 = \sum_i \sum_j (x_{ij} - \mu_i)^2. \tag{28}$$

Notice S is the within groups sum of squares as used in the ANOVA. Strictly speaking, one needs to determine K and n_i simultaneously. An obvious approach is to iterate over all possible combinations. This is of course not feasible. We introduce an effective heuristic here by showing a

numerical example. The idea is supposed by the following derivable system of equations.

$$\int_{a_{i-1}}^{a_i} [f(x)]^{1/3} dx = \int_{a_i}^{a_{i+1}} [f(x)]^{1/3} dx. \tag{29}$$

x	$f(x)$	$[f(x)]^{1/3}$	cum$[f(x)]^{1/3}$
$x = 10$	7	1.9113	1.9113
$x = 11$	6	1.817	3.73
$x = 12$	5	1.71	5.44
$x = 13$	5	1.71	7.15
$x = 14$	5	1.71	8.86

Given the above data, suppose we want to divide them into $K = 2$ groups. We first divide the last cumulative $[f(x)]^{1/3}$ by 2 and get 4.43. We then group $x = 10, 11$ together and $x = 12, 13, 14$ together.

6.2 Supervised Grouping

In this section, we present various methods of supervised grouping. Some of them utilizes the association measures we discussed early such as entropy and Z.

6.2.1 Baysian Average

Micci-Barreca(1998) presents a preprocessing scheme that converts categorical variable (unordered) with large cardinality into quasi-continuous scale variable. Applying a clustering on the subsequent continous variable is then equivalent to the grouping of the original categorical variable. The method is actually very simple. It is based on conditional probability estimate coupled with a Baysian average to count for small sample size.

Let y be a binary variable and x a high-cardinality categorical variable. One obvious transformation is the following:

$$x^i \to s_i = p(y = 1 | x = x^i).$$

The conditional probability can be easily estimated by $s_i = \frac{n_{i,y=1}}{n_i}$ where $n_{i,y=1}$ is the number of observations with $y = 1$ given $x = x^i$ and n_i is the number of observaions with $x = x^i$.

When the cardinality is high, some cell sizes may be very small and render the estimates unreliable. To mitigate the effect of small cells, s_i can be calculated as the mixture of two probabilities: the posterior probability of y given $x = x^i$ and the prior probability of y. Thus,

$$s_i = \lambda(n_i)\frac{n_{i,y=1}}{n_i} + (1 - \lambda(n_i))\frac{n_{y=1}}{n}.$$

$\lambda(m)$ is a monotonically increasing function between 0 and 1. The idea is that when n_i is large, the entire expression is dominated by posterior probability. When n_i is relatively small, we replace the probability estimate with the null hypothesis given by the prior probability of y. In other words, we in a sense assume that knowing $x = x^i$ does not help us in predicting y. Such a convex combination of prior and posterior probabilities is wided used in Emperical Bayesian.

There are many different choices for $\lambda(m)$. For our discussion, we should one of the functions used in tree algorithm C4.5

$$\lambda(m) = \frac{1}{1 + e^{\frac{n-k}{f}}}.$$

This is an s-Shaped function with parameter f controlling the slope of the function around the origin and k specfies the minimum cell sizes.

In the following, we give our SAS implementation of this technique.

```
/* Program: hwcatbavg.sas;
   Purpose: A macro pre-process categorical var;
   Idea: Categorical var with large cardinality
         can be transformed to a numerical variable;
         The summary or clustering of this num var
         can provide optimal grouping of the original cat var.
         The approach here is take the bayesian average of the
         posterior and prior prob lambda p(y=1|x) + (1-lambda) p(y=1)
         The lambda is used to adjust for small cells.
   Input: indat- input data set
          y    - binary or continuous;
          x    - the categorical var in question
                 it could be represented as numeric in SAS
          output- ouput dataset
```

```
          cluster- 1 if want a subsequent clustering analysis
   Author: Hongjie Wang
   Date: 11/15/2001;

   A note on Lambda: It is a monotonely decreasing function;
   The idea is that when lambda is large (when sample size is large),
   the posterior dominates;
   See Daniele Micci-Barreca's paper for the theory
*/


%macro hwcatbavg(indat,y,x,outdat,cluster);

%let K=500;
%let f=20;
proc sql;
   create table _hwcatp as
   select &x,
          count(*) as totx,
          avg(y) as meanyx
   from &indat(keep=&y &x)
   group by &x;

   select avg(y) into : meany
   from &indat(keep=&y &x);
quit;

data &outdat(drop=totx meanyx);
   set _hwcatp;
   lambda=1/(1+exp(-(totx-&K)/&f));
   prior=&meany;
   posterior=meanyx;
   score=lambda*posterior+(1-lambda)*prior;
run;

proc sort data=&outdat;
   by score;
run;
```

```
proc print data=&outdat;
run;


%if &cluster=1 %then %do;
   %sinclust(&outdat,score,&x);
%end;
%mend;


%macro sinclust(indat,score,idvar);
proc cluster data=&indat
     method=average std pseudo noeigen outtree=tree;
     id &idvar;
     var &score;
run;

proc tree;
run;
%mend;
```

We try this technique on our test data.

```
data test;
input x $  y c;
cards;
a1 0 148
a1 1 530
a2 0 111
a2 1 401
a3 0 645
a3 1 2229
a4 0 165
a4 1 890
a5 0 383
```

```
a5 1 2140
a6 0 96
a6 1 340
a7 0 98
a7 1 388
a8 0 199
a8 1 1029
a9 0 59
a9 1 229
a10 0 262
a10 1 1556
;
run;



data test(keep=x y);
   set test;
   do i=1 to c;
      output;
   end;
run;

%hwcatbavg(test,y,x,test,1);
```

The clustering structure coupled with business domain knowledge should provide good inputs for the grouping.

The technique can be easily extended to the case where y is continuous. In that case, instead of $p(y = 1|x = x^i)$, we are estimaing $E(y|x)$. Our SAS implementation handles both cases.

6.2.2 Crimcoord Approach

We discussed the technique Crimcoord in ??. This is the standard preprocessing procedure for categorical variable in decision tree software QUEST. It can be used to group a categorical variable. We apply this approach on our test data.

```
data test;
input x  y c;
cards;
1 0 148
1 1 530
2 0 111
2 1 401
3 0 645
3 1 2229
4 0 165
4 1 890
5 0 383
5 1 2140
6 0 96
6 1 340
7 0 98
7 1 388
8 0 199
8 1 1029
9 0 59
9 1 229
10 0 262
10 1 1556
;
run;



data test(keep=x y);
   set test;
   do i=1 to c;
      output;
   end;
run;

%crimcoord(test,y,x,test,x_num);

proc sort nodupkey data=test;
```

```
   by x;
run;

%macro sinclust(indat,score,idvar);
proc cluster data=&indat
     method=average std pseudo noeigen outtree=tree;
     id &idvar;
     var &score;
run;

proc tree;
run;
%mend;

%sinclust(test,x_num,x);
```

The results are fairly close to the one obtained from the Baysian average approach.

6.2.3 Correspondence Analysis Approach

Correspondence analysis is a very popular data reduction technique (among other applications) that is widely used in Europe and Japan. It transforms high dimensional data into a space of fewer dimensions. One of the purposes of CA is to present a graphical interpretation of data that may otherwise be difficult to understand. It is particularly powerful in dealing with contingency tables. CA transforms the table's categories into metrics, based on the table entries, and plots the categories in two or three dimensions. One can interpret distances between column items, between row items or between row and column items simultaneously. The number of dimensions extracted in CA equals $k-1$ where k=min(row dim, column dim). For a general discussion on correspondence analysis, see Greenacre (1993). An excellent discussion on using CA for contingency table data can be found in Kroonenberg and Lombardo (1999).

The following gives the SAS implementation. Notice the result again is very similar to the previous two techniques.

```
************************************************;
```

```
* program: hwcatcorresp.sas;
* purpose: grouping categorical variable;
* using correspondence analysis;
* Author: Hongjie Wang;
* Jan. 2002;
* %hwcatcorresp(indat,y,x,test,score);
* indat: input dataset;
* y: categorical outcome variable;
* x: categorical predictor;
* outdat: output dataset;
* score: the var name holding the numerical score;
************************************************;

%macro hwcatcorresp(indat,y,x,outdat,score);
data _hwcatcorr_temp1;
   set &indat(keep=&x &y);
run;
proc corresp data=_hwcatcorr_temp1 outc=&outdat short dim=1;
   table &y,&x;
run;

data &outdat(keep=&x &score);
   set &outdat(keep=dim1 _type_ _name_);
   if _type_='VAR';
   rename dim1=&score;
   rename _name_=&x;
run;
%mend;

%macro sinclust(indat,score,idvar);
proc cluster data=&indat
     method=average std pseudo noeigen outtree=tree;
     id &idvar;
     var &score;
run;

proc tree;
run;
```

```
%mend;



data test;
input x $  y c;
cards;
a1 0 148
a1 1 530
a2 0 111
a2 1 401
a3 0 645
a3 1 2229
a4 0 165
a4 1 890
a5 0 383
a5 1 2140
a6 0 96
a6 1 340
a7 0 98
a7 1 388
a8 0 199
a8 1 1029
a9 0 59
a9 1 229
a10 0 262
a10 1 1556
;
run;



data test(keep=x y);
   set test;
   do i=1 to c;
      output;
   end;
run;
```

```
%hwcatcorresp(test,y,x,test,score)
%sinclust(test,score,x);
```

6.2.4 χ^2 Merge Approach

χ^2 is tests the general association between two categorical variables. When we group x, we are essentially collapsing the columns of the underlying contingency table. The resulting χ^2 is likely to decrease. χ^2 merge approach combines different values of x in a way so that the reduction of χ^2 is minimized.

The likelihood ratio χ^2 has a very nice property. It can be partitioned. For example, suppose we have a two-way table where the column has values (A,B,C). It can be shown that $\chi^2((A,B,C)*y) = \chi^2((AB,C)*y)+\chi^2((A,B)*y)$. The following SAS program illustrates this interesting property. We point out that this property does not hold for the Pearson χ^2.

```
*****************************************************;
* Program: ChiPart.sas;
* Purpose: Demonstrate the Log Ratio Chisq Partition;
* Author: Hongjie Wang;
* The Log Ratio Chisq has a nice property;
* If one collaps A, B into AB;
* then the chisq of the A,B,C;
* equals chisq of A,B and AB, C;
* Such partition is not true for Pearson's Chisq;
* Such property is important for CHAID;
* where A,B will be merged if A,B not sig;
*****************************************************;

data ABC;
   input type $ response count;
   cards;
A 0 55
A 1 34
```

```
B 0 44
B 1 51
C 0 58
C 1 19
;
run;

proc sql;
   create table AB_C as
   select * from ABC
      where type='C'
   union
   select "AB" as type,
          response,
          sum(count) as count from ABC
      where type in ('A', 'B')
      group by response;
   create table AB as
   select * from ABC
      where type in ('A', 'B');
quit;

proc freq data=ABC;
   table response*type/chisq;
   weight count;
run;

proc freq data=AB_C;
   table response*type/chisq;
   weight count;
run;

proc freq data=AB;
   table response*type/chisq;
   weight count;
run;
```

The basic idea of χ^2 merge can be illustrated by the following example taken from Kass(1980).

The study covers 699 students who were admitted into different types of schools from 1960 to 1963. The goal is to assess any association between the year and the type of schools.

year	Others	JMB	TUEC	NSC
1960	7	39	107	17
1961	13	34	112	17
1962	11	30	93	26
1963	13	34	80	36

If we would calculate the subtable (year =1960, 1961), we see that the χ^2 is not significant.

```
data test;
   inputs type $ year count;
   cards;
other 1960 7
jmb   1960 39
tuec  1960 107
nsc   1960 17
other 1961 13
jmb   1961 34
tuec  1961 112
nsc   1961 17
other 1962 11
jmb   1962 30
tuec  1962 93
nsc   1962 26
other 1963 13
jmb   1963 34
tuec  1963 80
nsc   1963 36
;

proc freq data=test(where=(year<1962));
```

```
   table year*type/chisq;
   weight count;
run;
```

```
Statistics for Table of year by type
```

Statistic	DF	Value	Prob
Chi-Square	3	2.1532	0.5412
Likelihood Ratio Chi-Square	3	2.1809	0.5357
Mantel-Haenszel Chi-Square	1	0.3785	0.5384
Phi Coefficient		0.0789	
Contingency Coefficient		0.0786	
Cramer's V		0.0789	

Since χ^2 is testing the goodness-of-fit of an indepedent model, we conclude that knowing year 1960 or 1961 does not help us predict the type of schools better. Thus, not knowing the difference of these two years will not significantly reduce our ability to predict the outcome. Therefore, the first two rows can be merged together. The χ^2 merge algorithm simply keeping doing such kind of merge until all row-pair-wise χ^2 are significant. In CHAID, Bonferroni correction is done at each iteration to adjust for selection bias and a backtracking *unmerge* step is also included.

Notice the algorithm would require the forming of all possible pairs. It is computationally very expensive. Typically, some dynamic programming heuristics are applied. We skip the implementation in SAS at this point (It will be suited in a language that supports recursion.).

6.2.5 Zeta Maximization

In ?? we derived the Zeta statistics that can be used to measure association between two categorical variables. We can use this statistics to derive a discretization procedure.

Given a continuous x and a binary variable y, we want to create a binary indicator out of x (to deal with nonlinearity in x for example). The key is to find the "optimal" cutoff point.

7 Linear Regression

We cover several important topics in linear regression.

7.1 Nonlinearity and Nonadditivity

Linearity and additivity are two implicit assumptions of linear regression. Linearity refers to the requirement that y is related to x linearly, holding all the other variables in the model constant and such linear relationship remain the same for all the values of x. Additivity refers to the requirement that such linear relationship between y and x remains constant (the same slope), regardless of the values of other variables. In a sense, we can think of linearity as a special of additivity. It is trivial to realize that such assumptions are rarely true. The fact that multicollinearity is inevitable makes the statement "holding other variables constant" virtually impossible. Thus, we are interested in the degree of severity of nonlinearity and nonadditivity and their possible impact on the model.

It is important to point out that technically, nonlinearity and nonadditivity should be detected and deal with in a multivariate setting. Sometimes, it is not practical to do so, and bivariate screening is used to discover nonlinearity. While it is a useful and effective heuristic, we should use it with caution.

Consider the following example:

```
 data test;
   do i=1 to 100;
      x1=ranuni(10);
      x2=sin(x1)+rannor(20);
      y=2.5*x1+2*x2+rannor(30);
      output;
   end;
run;

%lowess(data=test,out=smooth, x=x1, y=y,
    htext=2,
    f=0.5, plot=YES,
    colors=blue red, outanno=lowess);
```

```
proc reg data=test;
   model y=x1 x2;
   plot residual.*x1;
run;
```

If we would examine the scatter plot of y versus x_1, we will notice a clear nonlinear relationship. But, this is not the technical definition of nonlinearity. In fact, if we see the way y is constructed, x_1 is expected to relate to y is a linear way holding x_2 constant. Here the problem is that the bivariate nonlinearity we observe between y and x_1 is really due to the presence of x_2 and its relationship with x_1. In fact, if we examine the residual plot, we see no evidence of any curvature.

The following is a non-graphic procedure to detect nonlinearity in a bivariate relationship.

Divide the entire population by x into three subsamples in the increasing order of x. Let z_1, z_2 be two indicators that classify observations into these subpopulations. Estimate the following model:

$$y = \alpha + \beta_1 x + \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 (x z_1) + \gamma_4 (x z_2) + \epsilon.$$

Nonlinearity is equivalent to $H_0 : \gamma_i = 0$, for $i = 1, 2, 3, 4$. An F test should do the trick in this case.

The non-additivity property can be similarly tested by dividing the population based on another predictor x'.

Introducing interaction term is one of the commonly used ways to handle nonadditivity. Lots of research argue that one should make sure the main effects are significant before adding any interaction terms. This approach tends to work less well in analyzing categorical variables. Consider the following example. Here we have two factors: gender (male=1) and package choice. The outcome variable is the purchasing indicator. We like to analyze the effects of gender and package (and possible interaction) on purchasing response rates.

```
data check;
   input gender package  buy count;
```

```
   cards;
1 1 1 29
1 1 0 171
1 0 1 46
1 0 0 160
0 1 1 17
0 1 0 73
0 0 1 12
0 0 0 77
;
data check(keep=gender package buy);
   set check;
   do i=1 to count;
      output;
   end;
run;

proc sort data=check;
   by gender;
run;


proc logistic descending data=check;
   model buy=package gender;
run;



proc logistic descending data=check;
   model buy=package;
   by gender;
run;
```

If we run the program, we would see that both main effects are not significant. But if we divide the population by gender and do a separate analysis, we observe that the package has a significant negative effect on men and a non-significant possible effect on women. These two effects with

opposite directions have similar size and end up offsetting each other.

A multiplicative model is in general more complicated to interpret than a simple additive model. In particular, the meaning of the coefficients and standard errors of the original variables change after interactive terms are introduced. For a thorough review of this subject, we refer to Friedrich (1982).

8 Logistic Regression Analysis

8.1 Derivations

Let X denote a matrix of K predictors and y be a binary variable. Furthermore, let $p_i = P(y_i = 1|X_i)$ for $i = 1, 2, \ldots, n$. The following expression gives the general linear model.

$$F^{-1}(p) = \beta X. \tag{30}$$

Here $F(t)$ is the cumulative probability function. When F is of logistic distribution, we have logistic regression. When F is normal, we have a probit model.

The logistic distribution is given by $F(t) = \frac{e^t}{1+e^t}$. Thus, the inverse is $log(\frac{p}{1-p})$. The logistic regression model is specified as

$$log(\frac{p}{1-p}) = \beta X \tag{31}$$

or equivalently

$$E(y|X_i) = p_i = P(y_i = 1|X) = \frac{e^{\beta_0+\beta X}}{1 + e^{\beta_0+\beta X}}. \tag{32}$$

The coefficients β is computed by maximum likelihood estimates. It can be shown that the MLE function is given by

$$logL(\beta) = \sum_{i=1}^{n} y_i \beta X_i - \sum_{i=1}^{n} log[1 + e^{\beta X_i}]. \tag{33}$$

Note: The ratio of probability of an event to the probability of an non-event is called Odds. Notice, Odds is not a linear function of X. As a result, logarithm transformation is applied to express this on an equal interval scale. The term *Logits* is a contraction of terms logistic and units.

8.2 Assess the Fit of Logistic Regression Models

In this section, we describe the Hosmer-Lemeshow test to assess the goodness of fit of logistic regression models.

Let p_{i1} denote the predicted probability for case i, $i = 1, 2, \ldots n$. The test is performed by first sorting the n observations by the predicted probabilities and forming M groups with approximately the same number of subjects in each group. This sorting strategy is usually called the 'decides of risk'. Let n_k be the number of subjects in the kth group, $k = 1, 2, \ldots, M$. The expected frequency of event $(Y = 1)$ is

$$E_{k1} = \sum_{i=1}^{n_k} p_{i1}. \tag{34}$$

The expected frequency of non-events $(Y = 0)$ is $E_{k0} = n_k - E_{k1}$. The actual observed frequency of events is $O_{k1} = \sum_{i=1}^{n_k} y_i$ and the observed frequency of non-events is $O_{k0} = n_k - O_{k1}$. The Hosmer-Lemeshow is given by

$$HL = \sum_{k=1}^{M} \sum j = 0^1 \frac{(O_{kj} - E_{kj})^2}{E_{kj}}. \tag{35}$$

The larger this statistics is, the more evidence of a lack of goodness of fit.(That is, if we see a large value with a small $p$value, we have evidence that the model is not good enough.)

Hosmer and Lemeshow showed that HL has an approximately Chi-Squared distribution with $M - 2$ degree of freedom. The use of $M - 2$ degrees of freedom was verified through extensive computer simulation. The reduced number of degrees of freedom is intended to account for the underdispersion in the test statistic caused by the varying p_ij terms in each group.

Problems will arise when the estimated probabilities are very small in certain cells. (The determinator is very small.) So, if we have a situation where all the predicted probabilities are close to 1 (which may be a good model), the test is likely to break down.

See *A cautionary note about assessing the fit of logistic regression models* by J. Pigeon and J. Heyse in Journal of Applied Statistics, Vol 26,no.7, 1999.

8.3 Log Likelihood Deviance Statistics

Given a saturated model, the MEL function is maximized with value 0(this is corresponding to a perfect fit.). The log-likelihood value of a fitted model can never be larger than the log-likelihood value for the saturated model. For a given model ,the deviance is defined as

$$DEV(X_0, X_1, \ldots, X_{p-1}) = 0 - 2logL(b_0, b_1, \ldots, b_{p-1}). \tag{36}$$

The smaller the deviance, the better the model is. This statistics can be used to compare two alternative models. In this case, the saturated model is used as the transitive benchmark.

In general, given two sets of variables $X_1, \ldots, X_q, X_{q+1}, \ldots, X_p$ and $X_1, \ldots, X_q$, the difference between $DEV(X_1, X_2, \ldots, X_p)$ and $DEV(X_1, X_2, \ldots, X_q)$ is approximately chi-square with $p - q$ degree of freedom. This allows us test the significance of additional variables. If a large deviance is detected, then, the variables $X_{q+1}, \ldots, X_p$ might be important to keep in the model.

8.4 Likelihood Ratio Test

It is equivalent to the deviance test. $(log(L1/L2) = log(L1) - log(L2).)$

8.5 Generalized R^2

Given a model with $X_1, X_2, \ldots, X_k$, L_0 denotes the likelihood function for the model containing only the intercept and L_M the likelihood function for the model containing all the predictors. The following expression defines the R^2. (There are lots of different versions of R^2, we only cover one that is implemented in SAS. For a complete discussion, we refer to (Menard 2000).)

R^2 is not a very reliable measure of predictiveness. Consider the following example. We have an almost perfect variable with no evidence of lack of fit, and yet the R^2 is ridiculously small. Also, this example illustrates the problem of using 0.5 as the optimal cutoff points.

The problem is due to the definition of R^2. It depends on the marginal distribution of x and y. The fact that y is very skewed and x is very uniform. For some alternative measures that are not marginal dependent, see Goodman (1991).

```
data check;
   inputs segment response count ;
   cards;
0 1 10
0 0 49990
1 1 1000
1 0 49000
;

data check(keep=segment response);
   set check;
   do i=1 to count;
      output;
   end;
run;
proc logistic data=check descending;
   model response=segment/rsquare ctable lackfit ;
run;
```

$$R_L^2 = \frac{-2[log(L_0) - Log(L_M)]}{-2[log(L_0)]}. \tag{37}$$

-2 log-likelihood statistics is similar to the sum of squared errors in the OLS. $-2[Log(L_0)]$ is similar to error variation of the model with only the intercept, analogous to the total sum of squares in OLS. Just like OLS tries to minimize the error sum of squares, the logistic regression tries to minimize -2 log-likelihood. Thus, R_L^2 can be interpreted as the proportional reduction in the -2 log-likelihood statistic. (This statistics is available directly from SAS by specifying the RSQ option in model statement.)

For a complete treatment on various R^2, see *Coefficients of Determination for Multiple Logistic Regression Analysis* by Scott Menard in *the American Statistician* February 2000.

Note: In OLS, the R^2 is related to the model F statistics in the following way, $F(k, n-k) = \frac{R^2/k}{(1-R^2)/(n-k)}$ where k is the number of variables and n is the number of cases.

8.6 Logistic Regression and Linear Discriminant Analysis

Let μ_1, μ_2 denote mean vectors for observations from groups A and B. If face a new case, the classification rule derived by linear discriminant analysis states if $(\mu_1 - \mu_2)'\Sigma^{-1}x_0 > (1/2)(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2)$ then x_0 goes to group A, otherwise goes to group B.

From the derivation of logisitic regression, one can see that under the idea situation of normal distribution, logistic regression yields the same rule as the LDA. This is important for two reasons

- Logistic regression is more general and can handle X where the distribution is not normal. It subsumes LDA in the special case of normal distribution.
- If the data is indeed normal, then the statistics calculated in LDA will be more powerful and efficient since it explicitly takes advantage of the assumption.

8.7 Ordered Logit Models

8.8 Multinomial Model

9 Discriminant Analysis

We give several alternative ways of deriving the multiple discriminant analysis model. Some insights can be obtained from this exercise.

9.1 Classification Approach

Consider the following

$$p(i|x) = \frac{p(i)p(x|i)}{\sum_i p(i)p(x|i)} \tag{38}$$

where $p(i)$ is the prior probability of coming from population i and $p(x|i)$ is the density of the p-dimensional vector x for population i. Notice the denominator of the above expression does not change and remains constant

for different i equations. Thus, $p(i|x)$ is really proportional to $p(i)p(x|i)$. If we take the log and let $C(i)$ denote the classification decision, then we have

$$C(i) = log(p(i)) + log(p(x|i)) \tag{39}$$

for population i. Notice we have assumed that $p(x|i)$ is normal for all i. Thus,

$$C(i) = log(p(i)) + log(\frac{1}{(2\pi)^{p/2}|V|^{1/2}} e^{-1/2(x-\mu_i)'V^{-1}(x-\mu_i)}) \tag{40}$$

or

$$C(i) = log(p(i)) + log(\frac{1}{(2\pi)^{p/2}|V|^{1/2}} - \frac{1}{2}(x-\mu_i)'V^{-1}(x-\mu_i)). \tag{41}$$

The above is the general quadratic discriminant function. Usually, we assume that V is the common to all population. If not, the above equation still holds. The classification is determined by selecting the largest $C(i)$. We can convert the above quadratic function into a linear form and yet keep the same classification result.

We consider the two class case for simplicity. But the multiple class case is similar.

Consider

$$C(1) - C(0) = x'V^{-1}(\mu_1 - \mu_0) - \frac{1}{2}(\mu_0 + \mu_1)'V^{-1}(\mu_1 - \mu_0) + log(\frac{p(1)}{p(0)}). \tag{42}$$

We are only interested in $C(1) - C(0) > 0$ or not. In this case, it solely depends on the result of evaluating $x'V^{-1}(\mu_1 - \mu_0)$. Notice, this time, the identical V for different populations is absolutely necessary. This is the classic Fisher's linear discriminant function. Notice, if we have m classes, $m-1$ functions are needed for classification. The cutoff point is determined by the constant term $-\frac{1}{2}(\mu_1 + \mu_0)'V^{-1}(\mu_1 - \mu_0) + log(\frac{p(1)}{p(0)})$.

It is worth emphasizing that the $L(x)$ is not probability. Rather, it is the difference of classification decisions.

9.2 Modeling Approach

Given a data matrix $(X_i|Y_i)$ for $i = 1, 2, \ldots, n$, where Y takes value 0 and 1 and X is a single explanatory variable, we consider the standard logistic regression model

$$\log(\frac{p(Y=1|x)}{p(Y=0|x)}) = \beta_0 + \beta_1 x. \tag{43}$$

The model assumes that the log ratio of the conditional probability density functions of Y given x is a linear function of x (here X is a random variable and x denotes a realization of X.). In the context of general additive model, the expression becomes

$$\log(\frac{p(Y=1|x)}{p(Y=0|x)}) = \beta_0 + \beta_1 g(x). \tag{44}$$

Here $g(.)$ is a measurable function that transforms x into another random variable.

In this note, we show that sometimes, the appropriate transformation $g(x)$ can be derived theoretically.

Let $f(x|y=k)$ denote the conditional pdf of X given $y = k(0,1)$, then the following relation holds true.

$$\log(\frac{p(Y=1|x)}{p(Y=0|x)}) = \log(\frac{f(x|y=1)}{f(x|y=0)}) - \log(\frac{p(Y=1)}{p(Y=0)}). \tag{45}$$

We give a proof of the above fact in the case where X is a discrete random variable. When X is continuous, the idea is similar.

Proof: Notice $\frac{p(Y=1|x)}{p(Y=0|x)} = \frac{p(Y=1,X=x)}{p(X=x)} \times \frac{p(X=x)}{p(Y=0,X=x)} = \frac{p(Y=1,X=x)}{p(Y=0,X=x)}$. Similarly, $\frac{p(x|Y=1)}{p(x|Y=0)} = \frac{p(X=x,Y=1)}{p(Y=1)} \times \frac{p(Y=0)}{p(X=x,Y=0)}$. The result follows by taking the log function from both sides.

This simple result is actually very useful. It establishes the fact that the logit model is a correct one if and only if the log ratio of the conditional densities of X given y is a linear function of x. The modeling process essentially tries to estimate $\frac{p(Y=1|x)}{p(Y=1|x)}$ given $\frac{p(x|Y=1)}{p(x|Y=0)}$.

If we know the closed form for $p(x|Y)$ in advance , we can derive the proper logit model.

The following result justifies linear discriminant analysis.

If $f(x|Y=k) \sim N(\mu_k, \sigma)$ where $k = 0, 1$, then $\log(\frac{p(Y=1|x)}{p(Y=0|x)})$ is a linear function of x.

Proof: Since $f(x|Y=k) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu_k)^2/2\sigma^2}$, it follows that $\frac{f(x|Y=1)}{f(x|Y=0)} = \exp[(x-\mu_0)^2/2\sigma^2 - (x-\mu_1)^2/2\sigma^2]$. Thus, $\log(\frac{f(x|Y=1)}{f(x|Y=0)}) = ((\mu_1-\mu_0)/\sigma^2)x + (\mu_0^2-\mu_1^2)/2\sigma^2$. The result follows directly from (3).

The key assumption here is the same variance in two sub-populations (Y=1 and Y=0), which is exactly what linear discriminant analysis assumes. In case the variance is not the same, Quadratic discriminant analysis is required. This exercise also shows when the normal condition is assumed, one can derive LDA from logistic regression. (Notice the α and β and the similar to the linear discriminant function and cutoff in LDA).

9.3 Projection and Optimization Approach

Let $a' = (\mu_1 - \mu_0)'V^{-1}$ and $y = a'x$. Then, we are effectively transforming a multi-normal distribution to a set of univariate distributions. It is desirable that if we can find the linear transformation such that ratio of between-group variance to within-group variance on the function y is maximized. (Notice, in the multiple class case, we again will have multiple functions and they should be orthogonal to each.) From geometric perspective, two centroids in the space of the original variables (p space) can be projected upon a single line and thus their difference can be characterized by a single dimension. In a three-group case, a plane can be passed through three points (three centroids) and therefore we need two dimensions (two discriminant functions) to describe that plane.

In essence, in the two group case, Fisher's linear discriminant function translates the two multivariate population into two univariate populations, and the two univariate population means are maximally separated relative to the within group population variance on the linear composition y. It can be shown that when the distribution condition is nice, it also has some optimality properties.

When the distribution is not known or not assumed, some nonparametric methods can be used to estimate $p(X|y)$. Two commonly used methods are kernel method and nearest neighbor method (both of them can be implemented in SAS).

Lots of mathematical programming based approaches are gaining lots of popularities. (see *Mathematical Programming Formulations for the Discriminant Problem: An Old Dog Does New Tricks* by Cliff Ragsdale.) These models can explicitly maximize the classification accuracy rate. One can

view MP-based classification as an extension of the ideas presented in Fisher's original derivation of the LDF, replacing Fisher's discrimination criterion by a classification accuracy criterion and replacing the linear function of LDF by one that is not necessarily linear.

Typically, MP-based approaches use the concept of boundary $f(b, x) = c$ where b, c are parameters. Thus, if $f(x_i, b) > c$ then i is in group 1. We can then view $|f(b, x_i) - c|$ as a heuristic index of confidence in the group assignment of entity. It represents external deviation when a misclassification occurs.

The practical problem with MP is that, because the MP method does not assume a distribution for the error terms, it produces estimates without any statistical properties. Hence, there is no way to evaluate whether the estimated parameters of MP models are statistically significant. However, it should be point out that in dealing with real data, assumptions are often violated and the calculated statistics may not be reliable to begin with.

There are also resampling estimation techniques (Jackknife, boostrap). For detail, see *development of Statistical Discriminant Mathematical Programming Model via Resampling Estimation Techniques* by H. Ziari, D. Leatham and P Ellinger in *American Journal of Agricultural Economics.*

9.4 Comparing DFA and Logistic Regression in Classification

Fan and Wang (1999) used a full crossed 3-factor experimental design (sample size, group proportions, and equal or unequal covariance) to compare these two methods. Their results show that in terms of classification error rates, these two methods are largely comparable. Logistic regression has more advantage when the covariances are not the same in two groups and therefore the assumptions of DFA are violated.

10 Clustering Analysis

10.1 Determine Number of Clusters

One Graphical method that might be useful is to plot the data using the first two principal components as dimensions. The first two components

usually are the most important ones in terms of the information they contain (variance is information).

11 Relationship between Chi-Square and Canonical Correlation

Given a a 2×2 contingency table with observed frequencies a, b, c, d for all four cells where $a + b + c + d = N$, it can be shown that $\chi^2(df = 1) = N\phi^2$, where $\phi = (ad - bc)/(\sqrt{(a+b)(c+d)(a+c)(b+d)})$. If we create a set of indicators for each category of the contingency table, we end up with two sets of indicators representing the row and column variable respectively. It is interesting that the canonical correlation analysis on these two sets of indicators are related to the Chi-Square analysis of the original contingency table. (See *Canonical Correlation and Chi-Square: Relationships and Interpretation* by W. P. Dunlap in *Journal of General Psychology* 2000.)

12 Data Reduction

The definition of the "best" strategy to produce a model which has good predictive power is difficult. There are lots of factors that will have impacts. Missing values have to be dealt with. Some statistical techniques handles missing values automatically, such as decision tree. Some requires explicit pre-processing. The problem with nonlinearity needs to be considered. Traditionally, such predictors are entered as linear terms or as dummy variables obtained after grouping. Categorization introduces problems of defining cutpoints, overparameterization and loss of efficiency. The problem of multicollinearity would complicate standard selection process such as stepwise regression. It is not clear what should be the optimal order according to which these pre-processing steps need to take place. But it is clear that their effects are not independent. Subject-matter knowledge should be used to guide selection and it is important to realize that a model which fits the current data set well may be too data driven to give adequate predictive accuracy in other settings. Moreover, the inclusion of unnecessary variables or over-complicated transformations has cost implications in that superfluous data will have to be collected and unjustified time have to be spent. Thus, the model builders should adopt a more pragmatic approach in which they

search, not for a true model, but rather for a parsimonious model which gives an adequate approximation to the data at hand.

Sauerbrei and Royston (1999) gives a framework in which variable selection is performed in a multivariate setting. Fractional polynomials are used to transform the predictors and cross-validation is used to check for the problem of overfitting. Chatfield (1995) illustrates the undesirable consequence of drawing inference on a model that is based on some kind of subset selection using the data of interest.As he puts it, when a model is formulated and fitted to the same data, inferences made from it will be biased and overoptimistic when they ignore the data analytic actions which preceded the inference. Miller (1984) authored the survey paper on subset selection of regression models in which he reviewed a comprehensive set of approaches prior to 1980.

12.1 Stepwise Regression

Incorrect Degrees of Freedom calculation. Stepwise techniques often capitalize greatly on even small amounts of sampling error and, thereby, reduce the generalizability of results. This in term leads to the problem of non-replicability. This problem is due to the greedy algorithms stepwise regressions typically use. Variables are entered one at a time within the context of previously entered variables, based on the additional variance explained. Suppose we have x_1, x_2 with infinitesimal difference. x_1 might be picked to enter even though x_2 might be more practical or its true population effect was even higher. When x_1 is selected, the algorithm would take a certain deterministic path.

The results of stepwise regression varies depends on the p value of accepting a variable and multicollinearity among predictors. One of the best research papers that give a comprehensive review of various algorithms, their weakness and situation where such application is appropriate is by Derksen and Keselman (1992).

12.2 Principal Component Analysis

Principal Component is one of most widely used multivariate techniques. It is typically used to summarize the data by transforming the original set of variables into smaller set of linear combinations that account for most of the variance of the original set.

12.2.1 Derivation

Let $X = x_1, x_2, \ldots, x_k$ be a data matrix of $n \times k$. Consider the following optimization problem:

$$\begin{array}{ll} \text{Minimize} & \text{Variance}(p_1) \\ \text{where} & p_1 = \sum_{i=1}^{k} w_{i,1} \times x_i \quad \text{and } w_1 * w_1' = 1. \end{array} \tag{46}$$

This is the first component. The rest of the components follows the same structure with the additional constraint that $p_{i+1} \times p_j = 0$ for $j = 1, 2, \ldots, i$.

It turns out that the solution to the above optimization problem is related to orthonormize the covariance matrix $X'X$.

Since $X'X$ is symmetric and full rank, it can be shown that there exists a $(k \times k)$ orthogonal matrix W, such that $W'W = WW' = I$ and $W'X'XW = \text{diag}(\lambda_1, \ldots, \lambda_k)$. Here $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$ are k eigenvalues of $X'X$ and P's columns consist of the corresponding egivenvectors. Let

$$P = XW = (p_1, p_2, \ldots, p_k) = (x_1, x_2, \ldots, x_k) \times \begin{pmatrix} w_{1,1} & w_{1,2} & \ldots & w_{1,k} \\ \ldots & \ldots & \ldots & \ldots \\ w_{k,1} & w_{k,2} & \ldots & w_{k,k} \end{pmatrix}.$$

Thus, $P'P = W'X'XW = \text{diag}(\lambda_1, \ldots, \lambda_k)$. p_i is the ith component and it is clear that $p_i \times p_j = 0$ for $i \neq j$ and $p_i \times p_i = \lambda_i$. The last expression shows that the variance of p_i is λ_i, for $i = 1, 2, \ldots, k$.

In most of the statistical softwares, the components are further normalized by $\Lambda^{1/2} = (\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_k})$ so that they are of unit length.

Here $w_{i,j}$ is the loading of variable x_i on the jth component. Some authors define the loading differently as the covariance of x_i and p_j. We follow the convention used by SAS and SPLUS. The confusion is understandable when we define loading in factor analysis. Given the definition we use, the loadings are the coefficients that are used in calculate linear combinations of the original variable to form the principal components.

We finish this section by giving an example. The following SAS program performs PCA manually using Singular Value Decomposition (SVD) and built-in SAS procedure *princomp*.

```
%let nvar=3;
%let n=10000;
```

```
  proc iml;
   R={1 0.5689 0.6130,
      0.5689 1 0.6882,
      0.6130 0.6882 1};
   means={0 0 0};
   sts={1 1 1};
    Z=shape(0,&n,&nvar);
    do n=1 to &n;
      do j=1 to &nvar;
        Z[n,j]=normal(0);
      end;
    end;
    call eigen(L,V,R); L=diag(L);
    X=(Z*sqrt(L)*V');
   do i=1 to &n;
      do j=1 to &nvar;
         X[i,j]=X[i,j]*sts[j]+means[j];
      end;
   end;
    varnames='x1':"x&nvar"; print varnames;
    create xdata from X[colname=varnames];
    append from X;
    close xdata;
    call eigen(L,V,R); L=diag(L);
print L;
print V;
proc princomp data=xdata  out=pscore; var x1 x2 x3; run;


Output from SVD. L contains eigenvalues (variance of PC)
V is the loading matrix.
            L

          2.2480523         0         0
                  0 0.4453246         0
                  0         0 0.3066231
```

```
                    V

    0.5565342 0.8146833 0.1629748
    0.5811098 -0.521893 0.6244512
    0.5937854 -0.252822 -0.763872


          Output from Proc princomp.


        Eigenvalues of the Correlation Matrix

      Eigenvalue    Difference    Proportion    Cumulative

1     2.22895267    1.77189546        0.7430        0.7430
2     0.45705720    0.14306707        0.1524        0.8953
3     0.31399013                      0.1047        1.0000


                    Eigenvectors

                Prin1         Prin2         Prin3

    x1       0.554609      0.823341      0.120492
    x2       0.583475      -.488032      0.649140
    x3       0.593267      -.289715      -.751066

    prin1=0.554609x1+0.583475*x2+0.593267*x3;
    0.554609 is not the correlation between x1 and p1!
```

12.2.2 PCA and Correlation

Since the input of PCA is the correlation or covariance matrix, we would expect some relationship. One of the most common confusion regards this question:

"Consider the component with the highest variance and variables that load high simultaneously on this component. Can we infer that there are serious multicollinearity among these variables?"

To answer to the above question is a surprising NO! PCA focuses on finding linear combinations that maximize the total variance in the data, not the covariance. However, PCA is intimately related to multicollinearity via its minor components (components with small variances).

If a component $p_z = \beta X$ has very small variance, then p_z is almost constant and close to 0 (since p_z is normalized). This implies that exists β such that $\beta X \approx 0$. This suggests that there are strong multicollinearity among variables with high coefficients. There are some elegant relationships that can be established between this property and variance inflation and condition index numbers. For detail, we refer to Hawkins and Fatti (1984).

12.2.3 Distance Invariance

Let $P = XW$ where $WW = I$. Then $P' = W'X'$ and $PP' = XWW'X' = XX'$ (This is not the covariance matrix, which is given by $X'X$.). Geometrically the rows of X give the co-ordinates of points in the original variable space and the rows of P gives co-ordinates relative to axes representing the new components. The invariance of the inner product matrices XX' and PP' ha sthe geometircal interpretation that Euclidean distances between individuals are preserved. That property gives justification for using PCA before distance-based clustering procedures. Also, mutlivariate outliers can be more easily detected by using PCA as the deviates will be summarized in the first couple of components.

12.2.4 PCA and Regression

We discuss two different issues here. First, we address how PCA can be used to deal with multicollinearity in regression. Second, we mention some PCA based variable selection approaches.

We start with the linear model $y = X\beta + \epsilon$ where ϵ is a vector of unobservable random disturbances which are independently and identically distributed normal random variables with $\mu = 0$ and variance θ^2. The least squares estimator of β is $b = (X'X)^{-1}y$ and under the assumption we have made, b has a multivariate normal distribution with a mean vector β and covariance matrix $\theta^2(X'X)^{-1}$. When there exists high level of multicollinearity, the variance of b tends to be very large and render the coefficient estimates reliable.

Perform a principal component analysis on X and we get $P = XW$ where

$WW' = W'W = I$. Thus, we have $y = X\beta + \epsilon = XWW'\beta + e = P\alpha + \epsilon$. The least squares estimate for α is $\hat{\alpha} = (P'P)^{-1}P'y$ and $b = W'^{-1}\hat{\alpha}$.

The chief use of the PCA is to detect and analyze multicollinearity in the original model and to provide a biased estimator of the original model but with smaller variance. In deriving calculating $\hat{\alpha}$, principal component regression amounts to dropping a component with the smallest eigenvalue. Notice, dropping a component is not equivalent to dropping variables. Rather it is like dropping some parts of all the original variables. Fomby showed that this PC regression has an interesting property. Dropping a component is equivalent to estimating b using LS by imposing one additional linear constraint to offset the effects of multicollinearity. As a result, the estimates are biased. It is also shown that PC regression is related to other biased approach such as root regression and ridge regression. It is important to point out that the PC regression still retains all the original variables. When replace the original regressor variables by their principal components, all the original variables have to be kept since any component consists of linear combination of all the original variables.

A slightly different approach is to retain only the original variables that contribute significantly to the major components. The idea is that variance is information. The minor components with little variance in general are not important in regression. Therefore, the variables that only load high on these components are not important either and can be discarded. Jellife (1972,1973) describes several methods that discard variables using PCA in a way that minimize the loss of information contained in the original dataset.

One of the methods tested to be effective works as follows:

1. Choose λ as the cutoff value below which a component would be considered information lacking.
2. Perform a PCA on all K original variables and order the resulting PCs $p_1, p_2, \ldots, p_K$ in decreasing order of their associated eigenvalues.
3. Let g be the index of the PC such that $\lambda(P_i) \leq \lambda$ for $i \geq g$.
4. For each x_i, calculate $weight_i = \sum_{j=g}^{K}(w_{i,j})^2$. Here $w_{i,j}$ is the coefficient of x_i on p_j.
5. discard $K - g$ variables with the highest weights.

After $K-g$ variables are discarded, one may optionally repeat the process again. There are several variants of this method detailed in Jolliffe's paper.

It is imperative that we point out these methods were originally designed to discard variables in the context of PCA without losing much information. The outcome variable y is not in the equation at all. However, they are extended to be used in regression variable selection by people since, after all, regression is to use the information in X to predict the information in y. In fact, this is one of the variable selection techniques described by SAS Institute in their Regression Course Notes.

This seemingly reasonable approach should be used with caution. First, the weights $w_{i,j}$ does not necessarily tell us the relative importance of variables in forming the principal components. This is similar to linear regression where the non-standarized coefficients do to correspond to importance of the variables. There are couple of ways to assess the importance of variables in forming a particular principal component. One of them is to compute the correlation between x_i and p_j. This is given by $w_{i,j}\sqrt{\frac{\lambda_j}{s_i}}$ where s_i is the variance of ith variable. Empirical study (see Cadmium and Jollies (1995))has shown that

1. Similar weights (coefficients), even very large ones, may translate into very different correlations between these variables and the PCs.

2. Other than the first component, where the largest weight corresponds to the highest correlation, the ranking of weights need not correspond to the ranking for correlations.

Secondly, it is possible that the first $K-g$ components contribute nothing to the reduction of sum square while the last component with the smallest eigenvalue explains all the covariance between X and y (see Had and Lan (1998)). Such a situation can occur whenever the β coefficient vector is in the direction of jth eightieth of $X'X$. In that case, only the jth component has predictive power for y. We give a SAS program to illustrate this phenomenon.

We create a data matrix of 4 variables. Some high multicollinearity is deliberate. The higher the collaterality, the more concentrated the variance will be by the first component. We then construct y as linear combination of X using the egienvector that corresponds to the last component as coefficients. When we regress y against the resulting principal components, we notice that only the last component is significant. Since the PCs are independent of each

other, dropping any one of them will not change the significance of others. Also, notice that the R^2 in the regression model exactly equals the square of the variance contained in the last component.

```
%let nvar=4;
%let n=10000;
  proc iml;

R={1.0   0.5  0.75  0.6,
   0.5   1.0  0.89   0.1,
   0.75  0.89 1.0    0.1,
    0.6  0.1  0.1    1.0};

   means={0 0 0 0};
   sts={1 1 1 1};
    Z=shape(0,&n,&nvar);
    do n=1 to &n;
      do j=1 to &nvar;
        Z[n,j]=normal(0);
      end;
    end;
    call eigen(L,V,R); L=diag(L);
    print L;
    print V;
    X=(Z*sqrt(L)*V');
   do i=1 to &n;
      do j=1 to &nvar;
         X[i,j]=X[i,j]*sts[j]+means[j];
      end;
   end;
    varnames='x1':"x&nvar"; print varnames;
    create xdata from X[colname=varnames];
    append from X;
    close xdata;
quit;

data xdata;
   set xdata;
```

```
   y=0.3-0.48614*x1-0.423471*x2+0.71768*x3+0.26318*x4+rannor(10);
run;

proc princomp data=xdata  out=scores;
   var x1 x2 x3 x4;
run;

proc reg data=scores;
   model y=prin1 prin2 prin3 prin4;
run;
```

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.53685815	1.37322939	0.6342	0.6342
2	1.16362876	0.86759229	0.2909	0.9251
3	0.29603647	0.29255985	0.0740	0.9991
4	0.00347662		0.0009	1.0000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4
x1	0.554040	0.313541	-.598579	-.486246
x2	0.527734	-.389695	0.626377	-.421054
x3	0.587038	-.309820	-.203560	0.719695
x4	0.264428	0.808605	0.455980	0.261378

The REG Procedure
Model: MODEL1
Dependent Variable: y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	46.80217	11.70054	11.84	<.0001
Error	9995	9879.02229	0.98840		
Corrected Total	9999	9925.82446			

Root MSE	0.99418	R-Square	0.0047
Dependent Mean	0.30452	Adj R-Sq	0.0043
Coeff Var	326.47432		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.30452	0.00994	30.63	<.0001
Prin1	1	0.00442	0.00624	0.71	0.4793
Prin2	1	0.01424	0.00922	1.54	0.1224
Prin3	1	0.02479	0.01827	1.36	0.1749
Prin4	1	1.10087	0.16862	6.53	<.0001

12.2.5 PCA of Categorical Variables

Technically speaking, PCA does not have any distribution constraints. However, it is traditionally applied on numerical (ordered) variables only. Applying PCA on categorical variables would require the construction of dummy variables like we do in linear regression. However, as a data summary or reduction method, it can be applied on categorical variables. In this section, we show a categorical variable pre-processing procedure called *crimcoord*. This procedure is implemented in decision tree algorithm QUEST (Quick Unbi-

ased Efficient Statistical Tree), which was developed by Loh and his students in University of Wisconsin, Madison. It is also commercially adopted in SPSS's Answer Tree.

Here is the outline of the algorithm.

Let y be the class variable and x be a categorical predictor with values $x^1, x^2, \ldots, x^K$.

1. Define $v_i = (x = x^i)$ for $i = 1, 2, \ldots, K$.
2. Perform a principal component analysis on $V = [v_1, v_2, \ldots, v_K]$.
3. Discard any components with little variance (close to zero).
4. Project the reduced data onto the largest discriminant coordinate.

The last step of the process can be accomplished by canonical discriminant analysis. In a canonical discriminant analysis we find linear combinations of the quantitative variables that provide maximal separation between the classes or groups.

The following SAS macro implements the algorithm.

```
   Program: crimcoord.sas;
   Author: Hongjie Wang;
   Purpose: Implement the crimcood procedure described in Quest ;
   See Advances in Decision Tree in KDD 2001 for some algorithmic details;
   Date: Jan. 2002;
   indat: input dataset;
   y: target variable (class var);
   x: categorical predictor;
   outdat: Ouput dataset;
   varname: Name of the variable holding numerical score;


%macro crimcoord(indat,y,x,outdat,varname);
data _hwcrim_temp1;
   set &indat(keep=&x &y);
run;
```

```
proc sql noprint;
   select distinct &x into : xvalues separated by ' '
   from _hwcrim_temp1;
   select count(distinct &x) into : x_cnt separated by ' '
   from _hwcrim_temp1;
quit;


data _hwcrim_temp1;
   set _hwcrim_temp1;
   %do I=1 %to &x_cnt;
      %let xvalue=%scan(&xvalues,&I, %str( ));
      xb&i=(&x="&xvalue");
   %end;
run;

proc princomp data=_hwcrim_temp1
     cov  noprint out=_hwcrim_temp2
          outstat=_hwcrim_temp3;
   var xb:;
run;

data _hwcrim_temp3(keep=xb:);
   set _hwcrim_temp3;
   if _TYPE_='EIGENVAL';
run;

proc transpose data=_hwcrim_temp3 out=_hwcrim_temp3;
run;

proc sql noprint;
   select count(*) into : npcs separated by ' '
   from _hwcrim_temp3
   where col1>=0.1;
quit;

PROC CANDISC data=_hwcrim_temp2 ncan=1
```

```
            out=&outdat(rename=(can1=&varname)) noprint;
   var prin1- prin&npcs;
   class &y;
run;

%mend;

/*
options mprint symbolgen;
data test;
input x $ y freq;
cards;
c1 1 4
c2 1 1
c3 1 5
c1 0 2
c2 0 2
c3 0 6
;
run;

data test(keep=x y);
   set test;
   do i=1 to freq;
      output;
   end;
run;


%crimcoord(test,y,x,test,x_num);

proc print data=test;
run;
*/
```

We shall compare this technique with some other categorical analysis

techniques in detail later on.

12.3 Factor Analysis

Common factor analysis was invented nearly 100 years ago by psychologist Charles Spearman. Its purpose is to discover simple patterns in the relationship among the variables. In particular, it seeks to investigate if the observed variables can be explained largely in terms of a much smaller number of variables called *factors*. It is widely used in the psychology and marketing research as a multivariate analysis technique and is a topic of countless research and application. For a general and comprehensive treatment, we refer to the unpublished manuscript *Exploratory Factor Analysis* by L. R. Tucker and R. C. MacCallum. In this section, we focus on some interesting properties and facts of factor analysis that are relevant to variable selection and modeling building.

12.3.1 Derivation

Before we get into the formal mathematical derivation, let's introduce the simple concept of partial correlation. Let x, y be two variables with correlation $r_{x,y}$. The correlation measures the linear relationship between these two variables. If we think of variance as information a variable contains, then correlation is the commonly shared information between these two variables. We further hypothesize that the one of the reasons x, y are related is due to the fact that they both are related to a third variable z. The partial correlation of x, y given z is defined as

$$r_{x,y|z} = \frac{r_{x,y} - r_{x,z}r_{y,z}}{\sqrt{(1 - r_{x,z}^2)(1 - r_{y,z}^2)}}.$$

The partial correlation measures the commonly shared information left between x, y after we consider their relationship with z. The left common variance can then possibly be explained by another variable w. Factor analysis is essentially a process to find such y and w.

We give an example to illustrate this concept. This is taken from Darlington(1995).

Consider x_1 to x_5 and the corresponding correlation matrix.

$$\begin{pmatrix} & x_1 & x_2 & x_3 & x_4 & x_5 \\ x_1 & 1.00 & 0.72 & 0.63 & 0.54 & 0.45 \\ x_2 & 0.72 & 1.00 & 0.56 & 0.48 & 0.4 \\ x_3 & 0.63 & 0.56 & 1.00 & 0.42 & 0.35 \\ x_4 & 0.54 & 0.48 & 0.42 & 1.00 & 0.3 \\ x_5 & 0.45 & 0.4 & 0.35 & 0.3 & 1.00 \end{pmatrix}.$$

The correlation matrix depicts the relationship among these 5 variables. It has a unique pattern that will allow us to find one factor such that given this factor, all the bivariate partial correlation becomes 0. We construct a factor variable g such as $r_{g,x} = (0.9, 0.8, 0.7, 0.6, 0.5)$. Using the partial correlation definition we introduced early, it is easy to verify that $r_{x_i,x_j|g} = 0$. Thus, all the common variances among x can be explained entirely by their relationship with g. This is the most parsimonious factor model we can find for this set of variables. Factor analysis is a technique to find such hidden factor g. The following SAS program verifies our claim. The result will show that the ML algorithm finds one factor that sufficiently explains all the significant covaraince (It fails to reject H_0 that the residual correlation matrix is a null matrix and therefore one factor is enough.).

```
proc iml;
R={1 0.72 0.63 0.54 0.45,
   0.72 1 0.56 0.48 0.4,
   0.63 0.56 1 0.42 0.35,
   0.54 0.48 0.42 1 0.3,
   0.45 0.4 0.35 0.3 1};
means={0 0 0 0 0};
sts={1 1 1 1 1};
 Z=shape(0,&n,&nvar);
 do n=1 to &n;
   do j=1 to &nvar;
     Z[n,j]=normal(0);
   end;
 end;
 call eigen(L,V,R); L=diag(L);
 X=(Z*sqrt(L)*V');
do i=1 to &n;
```

```
   do j=1 to &nvar;
      X[i,j]=X[i,j]*sts[j]+means[j];
   end;
end;
 varnames='x1':"x&nvar";
 create xdata from X[colname=varnames];
 append from X;
 close xdata;
 call eigen(L,V,R); L=diag(L);
 V=V*sqrt(L);
quit;


proc factor data=xdata method=ml priors=smc corr res;
run;
```

Now, we introduce the mathematical model for common factor analysis. We shall make no distinction between population and sample factor analysis and we shall focus on exploratory factor analysis.

The basic common factor analysis model can be expressed

$$X = \Lambda F + e$$

where $X = (x_1, x_2, \ldots, x_p), f = (f_1, f_2, \ldots, f_q), e = (e_1, e_2, \ldots, e_p)$ and Λ is a $p \times q$ matrix called **factor loadings**.

The model states that $x_i = \sum_{j=1}^{q} \lambda_{i,j} f_j + e_i$. Each variable can be written as a common part, which is a linear combination of the factors and a unique part in the form of an error term. Furthermore, the model requires that each error term is not correlated with each other or the factors. Thus, $E(ee') = \Psi = diag(\Psi_1, \Psi_2, \ldots, \Psi_p)$ and $\text{cov}(e, f') = 0$.

Let Σ_{xx} denote the covariance matrix of X, then it can be shown that

$$\Sigma_{xx} = \Lambda\Theta\Lambda' + \Psi.$$

Here Θ is a $q \times q$ symmetric covariance matrix of the factors. This is the decomposition of the covariance matrix of X into two parts. The part is a quadratic function consisting of the factor loadings and the factor covariance. The second part is the covariance matrix of the error terms.

If we further restrict that the factors are uncorrelated, then we have $\Sigma_{xx} = \Lambda\Lambda' + \Psi$.

It is important to point out that in general, the loadings are **not** correlations between variables and factors. Such a correlation matrix is called structure matrix. However, when factors are uncorrelated, then the pattern matrix is identical with the structure matrix. (Since $x_i = \sum_{j=1}^{q} \lambda_{i,j} f_j + e_i$, and $cov(f_i, f_j) = 0$ we have $cov(x_i, f_j) = \lambda_{i,j}$). Furthermore, in this case, $\sum_{j=1}^{q} \lambda_{i,j}^2$ is all the variance of x_i explained by the factors and $sum_{i=1}\lambda_{i,j}$ is the total variance contained in jth factor.

The existence of a solution to the factor model is obvious. We can always have $X = XI + 0$. The goal is to get a solution such that q is as small as possible and yet it fits the data reasonably well. Thus, getting the smallest residual covariance matrix is usually the optimization problem that drives the data fitting process. On the other hand, it is important to realize that given the same kind of fit, the solution is not unique either.

Suppose $X = \Lambda_1 F_1 + e_1$ be a solution, and A be an orthogonal matrix such that $AA' = A'A = I$. It is trivial to verify that $(\Lambda_1 A)(A' F_1) + e_1 = X$ is another solution with the same fit (error term).

12.3.2 Rotation

We demonstrated the fact that factor solution is not unique, given the same degree of goodness of fit or covariance explained. In fact, a factor solution can be rotated by any nonsingular matrix. Rotation can be formulated as an optimization problem. Let Λ_0, F_0 be the initial solution of a factor analysis, say, by principal component method. Rotation in essence is trying to find nonsingular matrix A such that $\Lambda_0 A, A^{-1} F_0$ is a new set factor solution with a more simple structure. There are lots of different simplicity criterion proposed by Cattell and Thurstone, among others.

There are two major rotation methods, *oblique rotation* and *orthogonal rotation*. Oblique rotation does not preserve the orthogonal structure of the factors but in general gives better fit. *Oblimin, Promax* are two commonly used oblique techniques. Orthogonal rotation on the other hand, requires the resulting factor pattern to be orthogonal. Commonly used orthogonal techniques include *Quartimax, Varimax*. For a nice treatment on this subject, we refer to Darton (1980).

The purpose of rotation is easy interpretation. As a result, the decision is subjective. For some guidance on factor pattern rotation and its subsequent

interpretation, see Merenda(1997) and Rothman(1996).

12.3.3 An Example

We use an example to illustrate various concepts we have introduced so far regarding factor analysis.

Consider the following SAS program. We construct the data in a way such that 2 factors are expected.

```
data xdata(drop=i);
   do i=1 to 1000;
      f1=ranuni(100);
      f2=rannor(0);
      x1=0.3*f1+4.6*f2+ranuni(200);
      x2=0.45*f1-0.67*f2+rannor(0);
      x3=-0.34*f1+0.92*f2+ranuni(300);
      x4=f1+0.45*f2;
      output;
   end;
run;

proc factor data=xdata method=principal rotate=varimax
   priors=smc corr res score outstat=fact;
   var x1 x2 x3 x4;
run;
```

The sample correlation matrix is given as follows:

	x1	x2	x3	x4
x_1	1	-0.54441	0.94356	0.86222
x_2	-0.54441	1	-0.53833	-0.3924
x_3	0.94356	-0.53833	1	0.75244
x_4	0.86222	-0.3924	0.75244	1

The factor pattern given by the initial principal factor solution is

	factor1	factor2	sum of sqrt
x_1	0.99031	0.01667	0.980991785
x_2	-0.53905	0.2005	0.330775153
x_3	0.94333	-0.15058	0.912545825
x_4	0.84712	0.27578	0.793666903

The first two columns are called factor loadings. Since we have requested an orthogonal factor solution, these loadings are actually the correlation between the factors and variables. The last column is called communality. They are the amount of variances of each original variable that are explained by the factors. Notice the communality is relatively small for x_2. This is due to the fact that x_2 has small correlations with other variables. Thus, it does not have lots of common variance to share with the others.

Taking the outer product of a column will give a correlation matrix of X explained by the factor corresponding to that column.

The following are two such matrices by two factors.

	x_1	x_2	x_3	x_4
x_1	0.980713896	-0.533826606	0.934189132	0.838911407
x_2	-0.533826606	0.290574903	-0.508502037	-0.456640036
x_3	0.934189132	-0.508502037	0.889871489	0.79911371
hline x_4	0.838911407	-0.456640036	0.79911371	0.717612294

	x_1	x_2	x_3	x_4
x_1	0.000277889	0.003342335	-0.002510169	0.004597253
x_2	0.003342335	0.04020025	-0.03019129	0.05529389
x_3	-0.002510169	-0.03019129	0.022674336	-0.041526952
x_4	0.004597253	0.05529389	-0.041526952	0.076054608

Adding these two matrices,we get the total variance/correlation of X explained by the factors.

	x_1	x_2	x_3	x_4
x_1	0.980991785	-0.530484271	0.931678964	0.84350866
x_2	-0.530484271	0.330775153	-0.538693327	-0.401346146
x_3	0.931678964	-0.538693327	0.912545825	0.757586757
x_4	0.84350866	-0.401346146	0.757586757	0.793666903

One of the conditions to check the adequacy of the model is to test H_0 that the residual correlation matrix is 0.

Notice in the SAS program, we have also requested a orthogonal rotation. The new factor pattern is of the following:

	factor1	factor2	sum of sqrt
x_1	0.72053	0.67958	0.980992457
x_2	-0.24586	-0.51993	0.330774345
x_3	0.57012	0.7665	0.912559064
x_4	0.79896	0.39411	0.793659774

Notice the communality remains the same. If we carry out the same exercise we did for the initial factor solution, we would discover that the total variances explained by the factors remain the same!.

12.3.4 Factor Analysis and PCA

The difference between factor analysis and PCA is very obvious. First, factor analysis assumes a probability model while PCA is a deterministic procedure. Second, factor scores are very different from component scores. Components are linear combination of the original variables while factors are the basis from which the original variables are linearly formed. Indeed, getting the factor score is not a trivial thing and requires regression like procedure, see Grice and Harris (1998). Third, PCA tries to find components that summarize the original data by explaining all the variance. Factor analysis on the other hand, attempts to find latent factors that will explain as much as possible the covariance among the original variables.

Despite the striking difference, there are lots of confusion between these two techniques. This is also complicated by the fact that PCA can provide an initial solution to factor analysis (sometimes called principal factor). In SAS, for example, *proc factor* can be used to perform both factor analysis.

Let $P = WX$ be the result of the PCA. Since $W'W = WW' = I$, we have $X = W'P$. Thus, W' provides an initial solution to the pattern matrix of a factor analysis with $e = 0$.

Let's consider the following example in SAS.

```
%let n=1000; %let nvar=5;
```

```
* Use IML *;
proc iml;
R={1 0.6 0.5 0.7 0.3,
   0.6 1 0.4 0.2 0.7,
   0.5 0.4 1 0.25 0.1,
   0.7 0.2 0.25 1 0.3,
   0.3 0.7 0.1 0.3 1};
means={0 0 0 0 0};
sts={1 1 1 1 1};
 Z=shape(0,&n,&nvar);
 do n=1 to &n;
   do j=1 to &nvar;
     Z[n,j]=normal(0);
   end;
 end;
 call eigen(L,V,R); L=diag(L);
 X=(Z*sqrt(L)*V`);
do i=1 to &n;
   do j=1 to &nvar;
      X[i,j]=X[i,j]*sts[j]+means[j];
   end;
end;
print L;
print V;
 varnames='x1':"x&nvar";
 create xdata from X[colname=varnames];
 append from X;
 close xdata;
quit;

proc princomp data=cordat type=corr;
run;

proc factor data=xdata m=principal nfactor=5;
run;
```

It is very important to realize that if we do not specify prior communality estimate, proc factor actually performs a PCA instead. Indeed, if we run the above program, we will see the pattern matrices from both procedures are the same after normalization by $\sqrt{\lambda}$. To perform a factor analysis, we need to specify *prior=smc*. By doing so, the correlation matrix's diagonal elements are replaced by the communalities.

Factor analysis can be conducted by Maximum likelihood method as well. In general, the solutions from ML tend to fit the data better, but it is computationally more expensive.

12.3.5 Factor Analysis based Variable Selection

Using factor analysis as a data reduction technique is well known in marketing research. The idea is that the underlying factors capture all the essential information and relationship in the original data. If the number of factors are smaller than the number of original variables, we can use the factor scores as the new variables. The problem with this approach is again the interpretation of the results. Also, factor analysis is a very data driven technique whose outcome requires lots of subjective inputs. Nevertheless, as an exploratory technique, it is effective in detecting the pattern of the relationships among variables and uncover the major underlying constructs of the data.

When the factors are orthogonal to each other, they represent latent independent dimensions of the data. We can reduce the number of variables by selecting the most important variables along each dimension. This is the place where proper rotation can play a critical role. Ideally, we want to have each factor (dimension) dominated by few key variables and each variable loads high on few factors. Unlike the pattern matrix in PCA, in factor analysis, if orthogonal factors are used, the entries of the pattern matrix stand for the correlation between the factor and the original variables.

We demonstrate this approach by using the example we gave in an early section. Here, x_1 is the variable that has the highest bivariate association with y. But as we discussed before, its predictiveness is due to its relationships with other variables and it should be discarded as part of the final model. We perform a factor analysis to see if we can get an insights on the pattern of the correlation matrix.

```
data data1;
```

```
input x1 x2 x3 x4 x5 y;
cards;
87 84 217   133   37 76
236   155   180   290   177   243
91 111   83 149   52 20
164   10 145   244   281   139
121   167   120   72 73 42
17 87 31 114   99 10
137   189   112   167   97 130
82 118   82 10 252   37
85 41 150   218   167   107
166   136   241   254   61 237
208   218   156   131   182   205
165   107   240   225   167   202
102   41 147   234   217   143
83 228   58 151   10 96
267   104   190   280   190   209
160   69 203   135   100   67
167   137   108   128   198   86
109   135   61 100   270   88
248   226   94 169   169   172
10 107   75 135   13 41
237   246   15 104   290   155
222   249   108   202   266   284
89 128   83 192   49 88
129   145   83 67 215   56
145   37 205   147   165   131
210   214   220   124   103   169
175   132   165   266   41 154
137   236   58 37 27 64
219   284   116   49 261   186
276   156   290   209   238   276
176   220   171   86 160   193
260   210   114   209   133   215
101   36 193   176   137   75
219   165   225   181   178   222
217   251   85 176   115   193
283   191   220   170   237   241
```

```
 35 187   86 58 45 40
 176   272   73 46 131   123
 68 194   116   65 23 60
 78 180   56 62 47 22
 100   103   80 113   180   58
 290   197   252   274   148   290
 157   129   172   200   53 148
 139   206   68 167   153   135
 155   249   111   70 177   174
 128   101   172   203   65 96
 153   197   183   190   20 154
 173   201   10 141   273   166
 134   156   140   91 95 97
 240   290   99 223   68 217
 ;

 proc factor data=data1 method=p priors=smc rotate=varimax corr res;
 var x1 x2 x3 x4 x5 y;
run;
```

Here is part of the outcomes.

Rotated Factor Pattern

	Factor1	Factor2	Factor3	Factor4
x1	0.58966	0.39055	0.38468	0.44041
x2	0.94820	-0.15568	-0.18160	-0.02913
x3	-0.08721	0.29406	0.88499	-0.00274
x4	-0.09246	0.90746	0.28157	0.01640
x5	0.03655	0.02795	-0.00339	0.92115
y	0.55887	0.55443	0.44653	0.37654

Notice we have included y variable in the factor analysis. Four factors are retained. From the loading matrix we see that y is associated almost equally with all four factors. x_1 displays the same pattern. This explains the fact that x_1 is related to all the other variables. However, we also notice that for each factor, x_1 is not the dominating variable. Using the procedure we described early, we would have discarded x_1 successfully.

12.4 Variable Clustering

One approach to variable reduction is a cluster analysis of variables. If we can somehow group variables into several clusters, then, we can possibly select handful variables from each cluster. Banks(1989) employed variable clustering to analyze the socioeconomic indexes associated with certain patterns of human-rights violation in the world.

The operation is very similar to the object clustering. To start, we need to define similarity measures. Instead of Euclidean distance, we use distance measures mutual covariation. By using covariation, we create groups of variables such that they are as correlated as possible within the cluster and as uncorrelated as possible with variables in other clusters. The following are some of the most commonly implemented in standard softwares.

- Let G_1, G_2 be two groups of variables (one variable is a special group). We define $d(G_1, G_2) = \max_{x \in G_1, y \in G_2,} |r_{x,y}|$, where $r_{x,y}$ is the correlation of x and y.
- $d(G_1, G_2) = (\sum_{x \in G_1} \sum_{y \in G_2} |r_{x,y}|)/(|G_1| + |G_2|)$.

The clustering algorithm follows the steps outlined below. Given K variables $x_1, x_2, \ldots, x_K$,

1. Let $G_i = \{x_i\}$ and define the $K \times K$ distance matrix using one of the measures suggested earlier.
2. Let G_t, G_s be two groups such that $d_(G_t, G_s)$ is the largest. Combine G_t, G_s into a new group G_v.
3. The iterative process continues until some stopping conditions are met.

Like other hierarchical clustering methods, typical stopping criterion include number of clusters formed or all the distances are below certain cutoff d_0.

Similar to factor analysis, we intend to select one variable out of each cluster. The advantage of clustering is that there is no overlap between clusters. Thus, in a sense, the variable clustering is like performing a factor analysis with a perfect rotation which results in a very simple structure. But which variable to select is itself a nontrivial topic. There are several heuristic proposed in the literature.

1. Choose one of the variables at random.
2. Choose one of the variables that enter the cluster the earliest. This is possible since we deal with hierarchical clustering.
3. Use statistics to weight the importance of a variable in relation to the cluster it belongs to. Select the variable with the highest weight from this cluster.
4. Choose the variable that has the highest bivariate association with y.

12.5 Sliced Inverse Regression

13 EM Algorithm

We attempt to give an very simple introduction on EM algorithms. For an excellent review, see Redner and Walker (1984).

First, let's briefly recall the definition of maximum-likelihood estimation problem. We have a density function $p(x|\theta)$ that is governed by the set of parameters θ (for example, p might be a set of Gaussians and θ could be the means and covariances). We also have a data set of size N, supposedly drawn from this distribution, $X = \{x_1, x_2, \ldots, x_N\}$. We therefore assume these random variables are IID. The resulting density for the sample is

$$p(X|\theta) = \prod_{i=1}^{N} p(x_i|\theta) = L(\theta|X). \tag{47}$$

Notice on the left hand side we have $p(X|\theta)$. Given the unknown (but existing) parameter, what is the probability of observing X. On the right hand side, given that we have observed X, what is the likely candidate for θ. The likelihood is thought of as a function of the parameters θ where the data X

is fixed. In the maximum likelihood problem, our goal is to find the θ that maximizes L. That is, we wish to find θ^* where

$$\theta^* = argmax_\theta L(\theta|X). \tag{48}$$

Often, we maximize $log(L)$ because it is analytically easier. Most of the time, an analytical solution is not available. An iterative numerical approach is taken.

The following exercise helps to illustrate the concept.

Suppose we have $p(x|\theta) = \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2}$, that is a normal distribution with mean μ and variance 1. We take $x_1, x_2, \ldots, x_n$ samples from this distribution. Then $log(L) = \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}}(-1)/2(x_i^- \mu)^2$. To maximize this expression is equivalent to minimizing $\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n} {x_i}^2 + n\mu^2 - (2\sum_{i=1}^{n} x_i)\mu$. For this quadratic form, the min occurs where $\mu = (\sum_{i=1}^{n} x_i)/n$. This is exactly the sample mean. Thus, the sample mean we use is also the maximum likelihood estimate for the population mean when the distribution is normal. (We can show the same for sample variance.) We also know that the sample mean formula is derivable from the Least Square approach (take a null model and the intercept is the sample mean). Thus, if X, y is normal, the least square solution in regression is consistent with the maximum likelihood solution. This may not be true for other distributions.

The EM algorithm is a general method of finding the maximum-likelihood estimate of the parameters of an underlying distribution from a given data set when the data is incomplete or has missing values.

There are two main applications of EM algorithm. The first occurs when the data indeed has missing values. The second occurs when optimizing the likelihood function is analytically intractable but when the likelihood function can be simplified by assuming the existence of and values for additional but missing (hidden) parameters.

Let $Z = (X, Y)$ where X is observed data and Y is the missing data. We also have to assume the existence of a joint density function:

$$p(z|\theta) = p(x, y|\theta) = p(y|x, \theta)p(x|\theta). \tag{49}$$

Here the joint density comes from two different sources. One, the missing values occur in the sample (that is, for a given variable, we have observed some values and missed some other values.). We assume a joint density between the observed and missed values. The second source comes from the situation where we have hidden parameters (segmentation membership, for example,

in the CRISP method where the response function and segmentation are derived at the same time.). We define

$$L(\theta|Z) = L(\theta|X,Y) = p(X,Y|\theta). \tag{50}$$

Notice, in the regular ML approach, we assume θ constant (we just have to find it my maximize L). In this case, we actually have a random variable, since Y is missing. Thus, we have $h_{X,\theta}(Y) = L(\theta|X,Y)$ where X, θ are constant and Y is a random variable. In the regular ML method, given a θ, we have a value for L. We just find the θ^* that maximizes L. In this case, given a θ, instead of getting a value of L, we get a function of Y. Suppose Y has a certain probability distribution, we can at least calculate the expected value of $h(Y)$, which is a number. Thus, we keep finding θ until we maximize the expected value of $h(Y)$. One can easily see the meaning of EM algorithm in this context.

Cadez *et al.,* (2000) developed a general framework for segmenting visitors on the web. Each customer is assumed to come from a segment. Within each segment, a Markov model is used as the generative process. The hidden segmentation membership and Markov models are estimated simultaneously by an EM algorithm.

References

[1] Avery, R. B, Bostic, R. W, Calem, and Canner, G. B. (2000), Credit Scoring: Statistical Issues and Evidence from Credit-Bureau Files, *Real Estate Economics*, Vol 28, No. 3, pp. 523-547.

[2] Banks, D. L. (1989), Patterns of Oppression: An Exploratory Analysis of Human-Rights Data, *Journal of American Statistical Association,* Vol. 84, Issue 407, pp. 674-681.

[3] Barnett, V. (1976), The Ordering of Multivariate Data, *Journal of the Royal Statistical Society, Series A,* Vol. 139, Issue 3, pp. 318-355.

[4] Bertrand, P. V. (1998), Constructing Explained and Explanatory Variables with Strange Statistical Analysis Results, *the Statisticians*, 47, pp. 377-383.

[5] Bhattacharyya, S. and Pendharkar, P. C. (1998), Inductive, Evolutionary, and Neural Computing Techniques for Discrimination: a Comparative Study, *Decision Sciences*, Vol 29, No. 4, pp. 871 - 899.

[6] Breiman, L. and Freedman, D. (1983), How Many Variables Should be Entered in a Regression Equation? *Journal of the American Statistical Association*, Vol. 78, No. 381, pp. 131-136.

[7] Bucklin, R. E. and Gupta, S. (1992), Brand Choice, Purchase Incidence, and Segmentation: an Integrated Modeling Approach, *Journal of Marketing Research*, Vol. XXIX, pp. 201-215.

[8] Cadez, I., Gaffney, S. and Smyth, P. (2000), A General Probabilistic Framework for Clustering Individuals, *Technical Report,* University of California, Irvine.

[9] Cadmium, J and Jollies, I. T. (1995), Loadings and Correlations in the Interpretation of Principal Components, *Journal of Applied Statistics*, Vol. 22, Issue 2, pp. 203-215.

[10] Chatfield, C. (1995), Model Uncertainty, Data Mining and Statistical Inference, *Journal of Royal Statistical Society A*, Vol. 158, pp. 419-466.

[11] Cooper, J. C. B. (1983), Factor Analysis: An Overview, *American Statistician*, Vol. 37, No. 2, pp. 141-147.

[12] Cooper, J. C. B. (1999), Artificial Neural Networks versus Multivariate Statistics: an Application from Economics, *Journal of Applied Statistics,* Vol. 26, No. 8, pp. 909-921.

[13] Cox, D. R and Snell, E. J. (1974), The Choice of Variables in Observational Studies, *Applied Statistics*, Vol 23, No. 1, pp 51 - 59.

[14] Darcy, R. and H. Aigner (1980), The Uses of Entropy in the Multivariate Analysis of Categorical Variables, *Journal of Political Science*, Vol. 24, No. 1, pp. 155- 174.

[15] Darlington, R. B. (1995), Factor Analysis, *Technical Report*, Cornell University.

[16] Darton, R. A. (1980), Rotation in Factor Analysis, *Statistician*, Vol. 29, Issue 3, pp. 167-194.

[17] Deal, K. and S. J. Edgett (1997), Determining Success Criteria for Financial Products: A Comparative Analysis of CART, Logit and Factor/Discriminant Analysis, *Service Industries Journal*, Vol. 17, No. 3, pp. 489-506.

[18] Derksen, S. and Keselman, H. J. (1992), Backward, Forward and Stepwise Automated Subset Selection Algorithms: Frequency of Obtaining Authentic and Noise Variables, *British Journal of Mathematical and Statistical Psychology*, 45, pp. 265-282.

[19] Dillon, W. R., Mulani, N. and Frederick, D. G. (1989), On the Use of Component Scores in the Presence of Group Structure, *Journal of Consumer Research*, Vol. 16, pp. 106-112.

[20] Dillon, W. R. and Goldstein, M. (1984), *Multivariate Analysis.* New York: Wiley.

[21] Drew, J. H., D. R. Mani, A. L. Betz and P. Datta (2001), Targeting Customers with Statistical and Data-Mining Techniques, *Journal of Service Research,* Vol. 3, No. 3, pp. 205-219.

[22] Dunlap, W. P. and Landis, R. S. (1998), Interpretations of Multiple Regression Borrowed from Factor analysis and Canonical Correlation, *Journal of General Psychology,* Vol. 125, No. 4, pp. 397-407.

[23] Editor (2001), Factor Analysis, *Journal of Consumer Psychology*, 10(1 & 2), pp. 75-82.

[24] Fan, X. and Wang, L. (1999), Comparing Linear Discriminant Function with Logistic Regression for the Two-Group Classification Problem, *Journal of Experimental Education*, Vol. 67, No. 3, pp. 265-286.

[25] Fayyad, U. M. (1992), On the Handling of Continuous-Valued Attributes in Decision Tree Generation, *Machine Learning*, 8. pp. 87-102.

[26] Fomby, T. B., R. C. Hill and S. R. Johnson (1978), An Optimal Property of Principal Components in the Context of Restricted Least Squares, *Journal of the American Statistical Association*, Vol. 73, Issue 361.

[27] Frank, K. A (2000), Impact of a Confounding Variable on a Regression Coefficient, *Sociological methods and Research*, Vol 29, No. 2, pp. 147-194.

[28] Friedman, J., T. Hastie and R. Tibshirani (1999), Additive Logistic Regression: a Statistical View of Boosting, *Technical Report, Stanford University.*

[29] Friedrich, R. J. (1982), In Defense of Multiplicative Terms in Multiple Regression Equations, *American Journal of Political Science,* Vol. 26, Issue 4, pp. 797-833.

[30] Gehrke, J. and Wei-Yin Loh (2001), Advances in Decision Tree Construction, *KDD, 2001.*

[31] Geman, S., Bienenstock, E. and Doursat, R. (1992), Neural Networks and the Bias-Variance Dilemma, *Neural Computation*, Vol 4. pp. 1- 58.

[32] George, E. I (2000), The Variable Selection Problem, *Journal of American Statistical Association*, Vol 95, No. 452.

[33] Glymour, C., D. Madigan, D. Pregibon and P. Smyth (1997), Statistical Themes and Lessons for Data Mining, *Data Mining and Knowledge Discovery*, Issue 1, pp. 11-28.

[34] Goodman, L. A. (1981). Criteria for determining whether certain categories in a cross-classification table should be combined with special reference to occupational categories in occupational mobility tables. *Amer. J. Sociol.* **87**, pp. 612-650.

[35] Green, P. E., F. J. Carmone and D. P. Wachspress (1977), On the Analysis of Qualitative Data in Marketing Research, *Journal of Marketing Research*, Vol. XIV, pp. 52-59.

[36] Grice, J. W. and Harris, R. J. (1998), A Comparison of Regression and Loading Weights for the Computations of Factor Scores, *Multivariate Behaviorial Research*, Vol. 33, No. 2, pp. 221-247.

[37] Gilula, Zvi and S. J. Haberman (1986). Canonical Analysis of Contingency Tables by Maximum Likelihood, *Journal of the American Statistical Association*, **395**, pp. 780-788.

[38] Hall, M. (1998), Feature Selection for Discrete and Numeric Class Machine Learning, *Technical Report*, University of Waikato.

[39] Hawkins, D. M and Fatti, L. P. (1984), Exploring Multivariate Data Using the Minor Principal Components, *Statistician*, Vol. 33, Issue 4, pp. 325-338.

[40] Hirji, K. K (2001), Exploring Data Mining Implementation, *Communications of the ACM*, Vol. 44, No. 7, pp. 87-93.

[41] Hodges, S. D. and P. G. Moore (1971), Data Uncertainties and Least Squares Regression, *Applied Statistics.*

[42] Hosking, J. R. M., E. P. D. Pednault and M. Sudan (1997), A Statistical Perspective on Data Mining, *Research Report, IBM Research Division.*

[43] Jollies, I. J. (1972), Discarding Variables in a Principal Component Analysis, *Applied Statistics.*

[44] Kass, G. V. (1980), An Exploratory Technique for Investigating Large Quantities of Categorical Data, *Applied Statistics,* Vol. 29, Issue 2, pp. 119-127.

[45] Kerosene, I. and Hong, S. J. (1997), Attribute Selection for Modeling, *IBM Wasting Research Center Technical Report.*

[46] Kim, H. and W. Y. Loh (2001), Classification Trees with Unbiased Multiway Splits, *Journal of American Statistical Association*, Vol . 96, No. 454, pp. 589-604.

[47] , A. M. and P. E. Green (1996), Modifying Cluster-Based Segments to Enhance Agreement with an Exigency Response Variable, *Journal of Marketing Research*, Vol., pp. 351-363.

[48] , W. J. (1987), Selection of Variables to Preserve Multivariate Data Structure, Using Principal Components, *Applied Statistics*, Vol. 36, No. 1, pp. 22-33.

[49] Lamely, D. N. and Maxwell, A. E. (1973), Regression and Factor Analysis, *Biomedical*, Vol. 60, Issue 2, pp. 331-338.

[50] Lipstick, S. R. and M. Parent (1995), Sample Size Calculations for Non-Randomized Studies, *The Statistician*, Vol. 44, No. 1, pp. 81-90.

[51] Loh, Wei-Yin and Yu-Shan Shih (1997), Split Selection Methods for Classification Trees, *Statistica Sinica*, Vol. 7, pp. 815-840.

[52] Malunion, J. G. (1992), Model Specification Tests and Artificial Regressions, *Journal of Economic Literature* Vol. , pp. 102- 146.

[53] Madonna, V., Jail, A. K. and Beamier, M. (1977), Parameter Estimation in Marketing Models in the Presence of Multicollinearity: an Application of Ridge Regression, *Journal of Marketing Research.*

[54] Magidson, J. (1981), Qualitative Variance, Entropy, and Correlation Ratios for Nominal Dependent Variables, *Social Science Research*, 10, pp. 177-194.

[55] Manias, M. L. and Weaker, W. E. (1998), Correcting for Omitted Variables and Measurement Error Bias in Regression with an Application to the Effect of Lead on, *Journal of American Statistical Association*, Vol. 93, No. 442, pp. 494-504.

[56] Mason, C. H. and Permeable,., W. D. (1991), Collaterality, Power, and Interpretation of Multiple Regression Analysis, *Journal of Marketing Research.*

[57] mcdonald, P. (1999), What is a Statistical Model? *Technical Report, Department of Statistics, University of Chicago.*

[58] Menard, S. (2000), Coefficients of Determination for Multiple Logistic Regression Analysis, *The American Statistician*, Vol. 54, No. 1, pp. 17-24.

[59] Merenda, P. F. (1997), A Guide to the Proper Use of Factor Analysis in the Conduct and Reporting of Research: Pitfalls to Avoid, *Measurement and Evaluation in Counseling and Development.*

[60] Micci-Barreca(2001), A Preprocessing Scheme for High-Cardinality Categorical Attributes in Classification and Prediction Problems, *SIGKKD Explorations.*

[61] Miller, A (1984), Selection of Subsets of Regression Variables, *Journal of Royal Statistical Society A*, Vol. 147, pp. 389-425.

[62] Peacock, P. R. (1998), Data Mining in Marketing: Part 1, 2, *Marketing Management.*

[63] Pickering, J. F. and B. C. Isherwood (1975), Determinants of Expenditure on Consumer Durables, *Journal of the Royal Statistical Society, A*, Vol. 138, Issue 4, pp. 504-530.

[64] Pregibon, D and Vardi, Y. (1985), Estimating Optimal Transformations for Multiple Regression and Correlation: Comment, *Journal of American Statistical Association*, Vol 80, Issue 391.

[65] Quester, P. and E. Dion (1997), Scaling Numerical Variables and Information Loss: An Appraisal of Morrison's Work, *Marketing Bulletin*, Vol. 8, pp. 59-66.

[66] Ramsey, J. B. (1969), Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis, *Journal of the Royal Statistical Society (B),* Vol. 31, Issue 2, pp. 350-371.

[67] Rao, P. and R. L. Miller (1971), Applied Econometircs. Belmont, CA: Wadsworth.

[68] Redner, R. A. and Walker, H. F. (1984), Mixture Densities, Maximum Likelihood and the EM Algorithm, *SIAM Review*, Vol. 26, Issue 2, pp. 195-239.

[69] Ribic, C. A and Miller, T. W. (1998), Evaluation of Alternative Model Selection Criteria in the Analysis of Unimodal Response Curves using CART, *Journal of Applied Statistics* Vol. **25**, pp 685-698.

[70] Rivals, I. and L. Personnaz (1999), On Cross Validation for Model Selection, *Neural Computation,* Vol. 11, pp. 863-870.

[71] Rodgers, J. L (1999), The Bootstrap, the Jackknife, and the Randomization Test: A Sampling Taxonomy, *Multivariate Behavioral Research*, Vol. 34, No. 4, pp. 441-456.

[72] Rothman, J. (1996), Some Considerations Affecting the Use of Factor Analysis in Market Research, *Journal of the Market Research Society*, Vol. 10, No. 3, pp. 371-381.

[73] Sawiris, M. (1998), Optimum Grouping and Boundary Problem, *Journal of Applied Statistics*, Vol. 27.

[74] Sauerbrei, W. and Royston, P. (1999), Building Multivariable Prognostic and Diagnostic Models: Transformation of the Predictors by Using Fractional Polynomials, *Journal of Royal Statistical Society A*, Vol. 162, pp. 71-94.

[75] Schoenbachler, D. D., G. L. Gordon, D. Foley and L. Spellman (1997), Understanding Consumer Database Marketing, *Journal of Consumer Marketing*, Vol. 14, No. 1, pp.5-19.

[76] SHarma, S. and James, W. L. (1981), Latent Root Regression: an ALternate Procedure for Estimating Parameters in the Presence of Multicollinearity, *Journal of Marketing Research.*

[77] Shaw, M. J., C. Subramaniam, G. W. Tan and M. E. Welge (2000), Knowledge Management and Data Mining for Marketing, *Decision Support Systems*, 31, pp. 127-137.

[78] Siciliano, R. and Mola, F. (2000), Multivariate Data Analysis and Modeling through Classification and Regression Trees, *Computational Statistics & Data Analysis*, Vol. 32, pp. 285-301.

[79] Sloane, D. and S. P. Morgan (1996), An Introduction to Categorical Data Analysis, *Annual Review of Sociology*, No. 22, PP. 351-375.

[80] Soofi, E., Retzer, J. and Yasai-Ardekani, M. (2000), A Framework for Measuring the Importance of Variables with Application to Management Research and Decision Models, *Decision Sciences*, Vol. 31, No. 3, pp. 595-625.

[81] Stewart, G. W (1987), Collaterality and Least Squares Regression, *Statistical Science*, Vol. 2, Issue 1, pp. 68-84.

[82] Tipping, M. E. and Bishop, C. M (1999), Probabilistic Principal Component Analysis, *Journal of Royal Statistical Society B*, 61, part 3, pp. 611-622.

[83] Whitaker, J. S. (1997), Use of Stepwise Methodology in Discriminant Analysis, *Technical Report*, Texas A & M University.

[84] White, H. (2000), A Reality Check for Data Snooping, *Econometrica*, Vol. 68, No. 5, pp. 1097-1126.

[85] Woolley, k. (1997), How Variables Uncorrelated with the Dependent Variable Can Actually Make Excellent Predictors: the Important Suppressor Variable Case, *Southwest Educational Research Association.*

[86] Yin, P. and Fan X. (2001), Estimating R^2 Shrinkage in Multiple Regression: a Comparison of Different Analytical Methods, *Journal of Experimental Education*, Vol. 69, No. 2, pp. 203-224.

[87] Zhang, H. (1998), Classification Trees for Multiple Binary Responses, *Journal of American Statistical Association*, Vol. 93, No. 441, pp. 180-193.

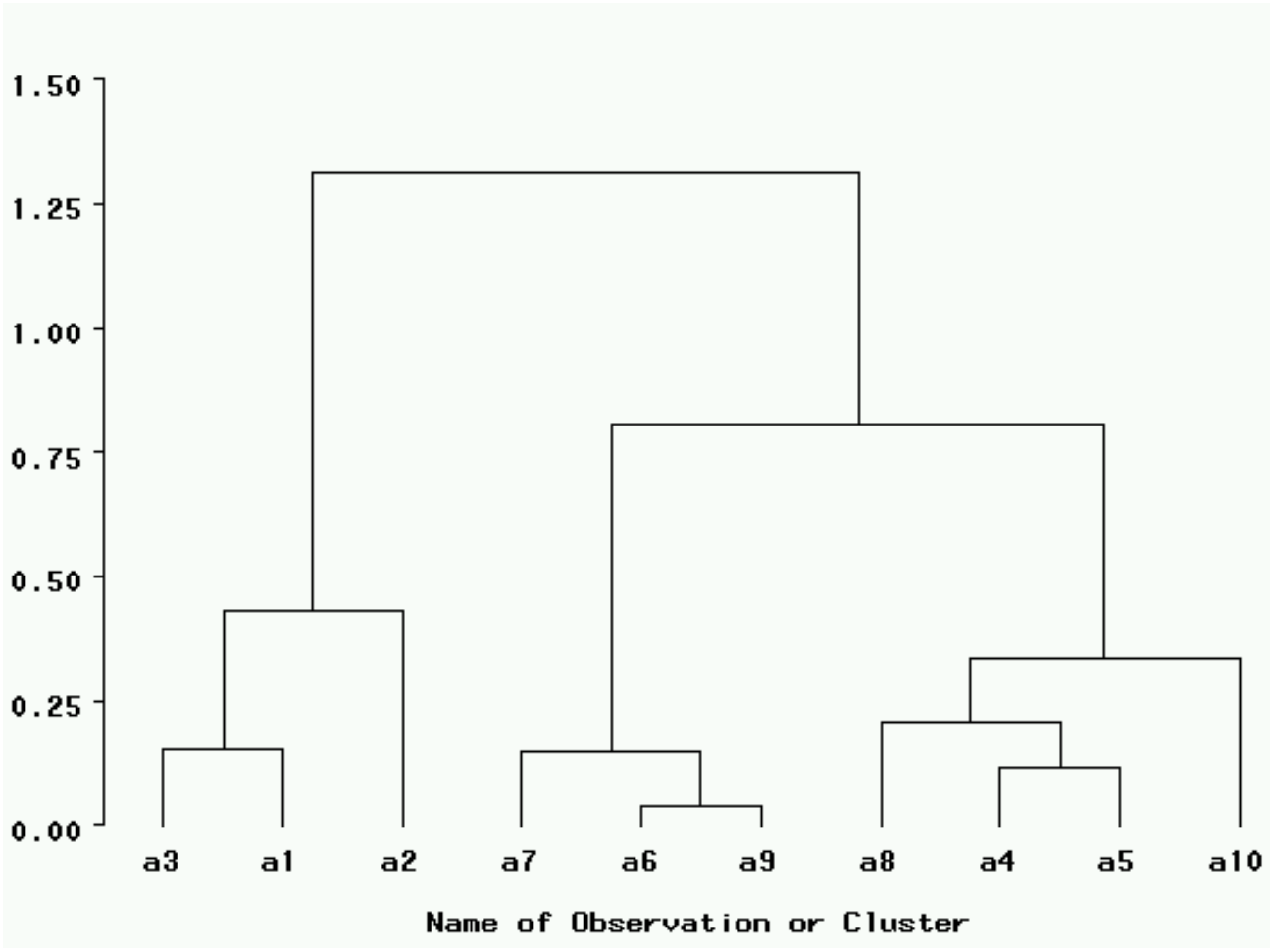

Figure 1: Baysican Average Approach

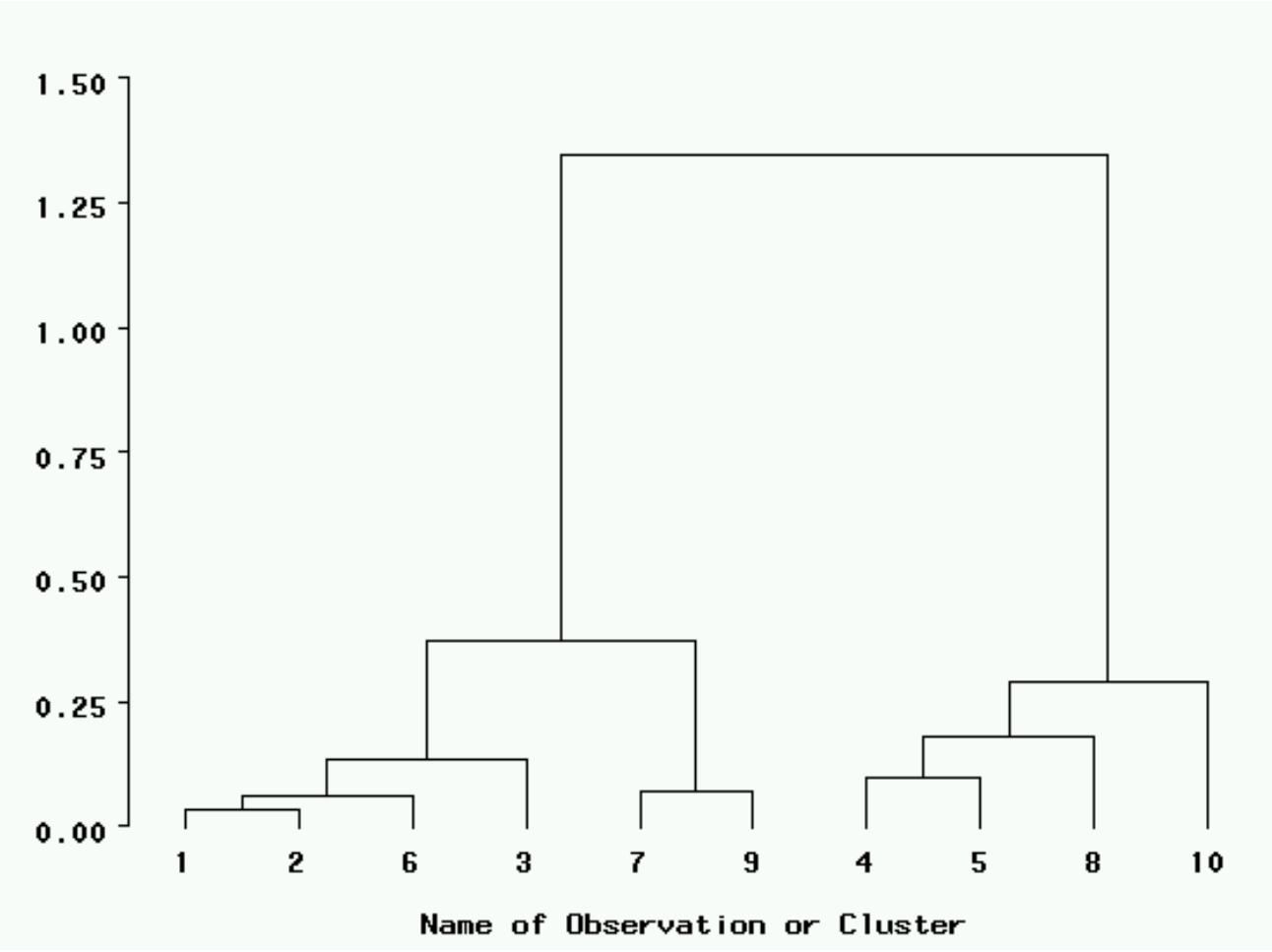

Figure 2: Crimcoord Approach

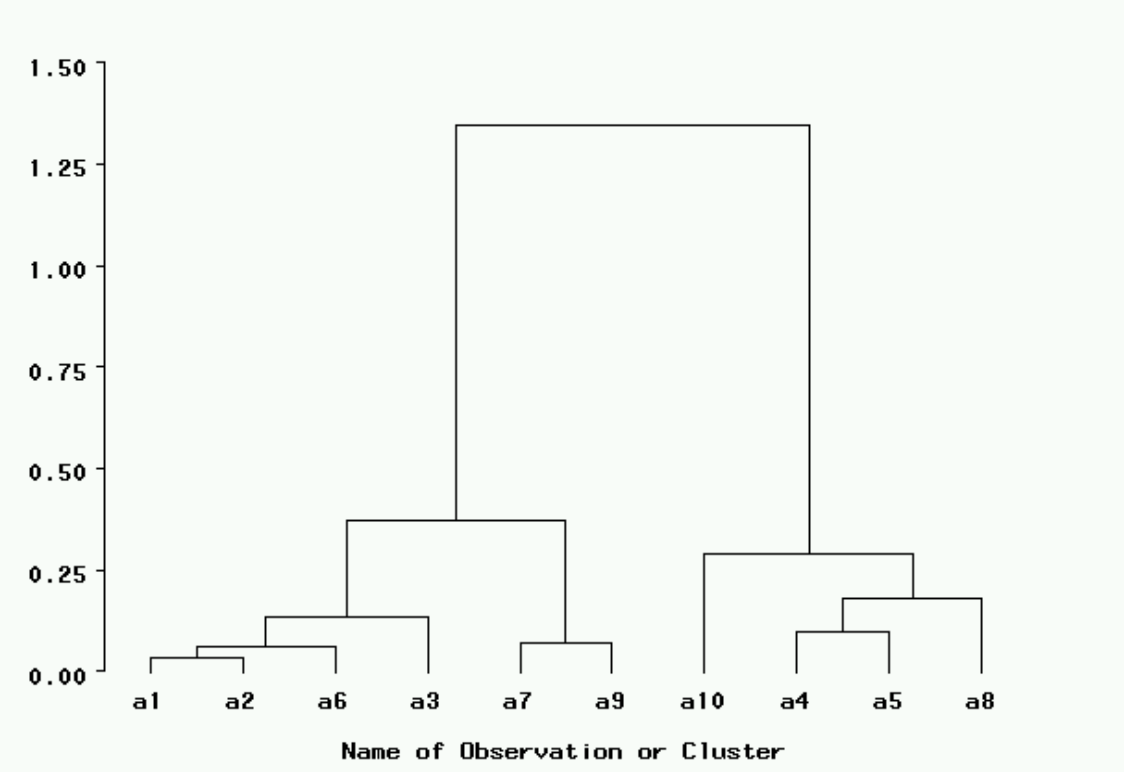

Figure 3: Correspondence Analysis Approach