Non-Parametric Analysis of Binary Choice Data

Keith T. Poole
Graduate School of Industrial Administration
Carnegie-Mellon University

16 June 1997

I thank Nolan McCarty, Howard Rosenthal, and Jerry Salancik for many helpful comments.

Abstract

This paper shows a general non-parametric technique for maximizing the correct classification of binary choice or two-category data. Two general classes of data are analyzed. The first consists of binary choice matrices such as congressional roll calls or preferential rank ordering of stimuli gathered from individuals. For this class of data a general non-parametric unfolding procedure is developed. To unfold binary choice data two subproblems must be solved. First, given a set of chooser or legislator points a cutting plane through the space for the binary choice must be found such that it divides the legislators into two sets that reproduce the actual choices as closely as possible. Second, given a set of cutting planes for the binary choices a point for each chooser or legislator must be found which reproduces the actual choices as closely as possible. Solutions for these two problems are shown in this paper.

The second class of data analyzed consists of a two-category dependent variable and a set of independent variables. This class of data is a subset of the binary choice unfolding problem. The cutting plane procedure can be used to estimate a cutting plane through the space of the independent variables that maximizes the number of correct classifications. The normal vector to this cutting plane closely corresponds to the beta vector from a standard probit, logit, or linear probability analysis.

1. Introduction

Categorical choice data occur frequently in the social and behavioral sciences. The purpose of this paper is to show a general non-parametric technique for maximizing the correct classification of binary choice or two-category data. I analyze two general classes of such data. The first class consists of binary choice matrices such as congressional roll calls or preferential rank ordering of stimuli gathered from individuals. For this class of data a general non-parametric unfolding procedure is developed. To unfold binary choice data two subproblems must be solved. First, given a set of chooser or legislator points, a cutting plane through the space for the binary choice must be found such that it divides the legislators into two sets so that the number of correct classifications of the legislators is maximized. Second, given a set of cutting planes for the binary choices, a point for each chooser or legislator must be found such that the number of correct classifications of the choices is maximized. Sections 3 and 4 below show solutions for these two problems.

The second class consists of data sets that would normally be analyzed with probit, logit, or linear probability models; that is, data sets with a two-category dependent variable and a set of independent variables. The cutting plane procedure developed in section 3 can be used to estimate a cutting plane through the space of the independent variables such that the number of correct classifications of the dependent variable is maximized. The counterpart to the beta vector from a probit, logit, or linear probability model is the normal vector of this cutting plane. In a Monte-Carlo analysis below I show that if the underlying error process is symmetric, the normal vector which maximizes classification and the beta vector from a probit analysis are virtually identical.

I make only two assumptions: 1) the choice space is Euclidean; and 2) the individuals making choices behave as if they utilize symmetric, single-peaked preferences.

The techniques I develop below are similar in spirit to those pioneered by Shepard (1962a,b) and Kruskal (1964a,b). They developed what became to be known as non-metric multidimensional scaling. Their initial focus was on similarities data. Their idea was to reproduce the rank ordering of the data as "closely" as possible. This became known as a non-metric approach in contrast to a metric approach such as factor analysis which treated the similarities as ratio scale data. The idea was to estimate a set of points in a Euclidean space such that the interpoint distances between the stimulus pairs were in the same rank order as the observed similarities.

Instead of placing points in a Euclidean space in such a way so as to reproduce the rank ordering of a set of data, I estimate cutting planes or cutting planes and chooser points in an Euclidean space such that the correct classification of the observed two-category data is maximized.

The paper proceeds as follows: section 2 defines the problem and explains the notation I use throughout the paper; section 3 develops the cutting plane procedure; section 4 shows how to estimate the legislator/chooser points; section 5 shows Monte-Carlo results for the roll call problem; section 6 shows empirical applications; and section 7 concludes.

2. Notation and Definitions

I begin by defining the problem in terms of the first class of data – a binary choice matrix – because the second class of data – a set of independent variables and a two-category dependent variable – is a special case of the binary choice matrix problem.

Given a matrix of binary choice data, the problem consists of finding a set of legislator points and a set of cutting planes corresponding to each binary choice in an Euclidean space of s dimensions such that each cutting plane divides the legislators into two sets that reproduce the actual choices as closely as possible. In other words, the task is to estimate a set points for the legislators (subjects) and a set of cutting planes for the roll calls (stimuli) that maximize the correct classification of the observed choices.

Let i=1,...,p be the number of legislators, j=1,...,q be the number binary choices (hereafter referred to as roll calls), k=1,...,s be the number of dimensions, **X** be the p by s matrix of legislator coordinates, and let **T** be the p by q matrix of observed choices. The choices will simply be yea or nay which I will represent as "y" and "n" respectively. **T** can contain missing entries.

For data sets which consist of a set of independent variables and a two-category dependent variable, the notation above is simply redefined. Let **X** be the p by s matrix of independent variables, where p is the number of observations and s is the number of independent variables, and let **t** be the p length vector which is the two-category dependent variable. In the development of the cutting plane procedure below, I will use the notation defined for a matrix of binary choice data. I will return to the problem of a set of independent variables and a two-category dependent variable in section 3.c.

Technically, a plane is defined as $\mathbf{z}'\mathbf{n} = \mathbf{v}'\mathbf{n}$ where **z**, **n**, and **v** are s by 1 vectors and the plane consists of all points **z** such that (**z** - **v**) is perpendicular to the normal vector, **n**, and **v** is some point in the plane. In simple algebra, an example of a plane in three dimensions is

$$3z_1 - 2z_2 + z_3 = 5$$

This plane is perpendicular to the normal vector -- $\underline{\mathbf{n}}'$ = (3, -2, 1) -- and $\underline{\mathbf{v}}$ is any point in the plane, for example (1, -1, 0), such that $\underline{\mathbf{v}}'\underline{\mathbf{n}}$ = 5.

In this context, the problem is to solve for the normal vector, $\underline{\mathbf{n}}$. Let **N** be the q by s matrix of normal vectors for the q cutting planes. Given the number of dimensions, s, the classification problem consists of finding estimates of **X** and **N**, which I will denote as **X*** and **N*** respectively, which maximize the correct classifications.

Maximizing the correct classification of a binary choice matrix can be broken down into two subproblems – 1) given the legislator coordinates, find the optimal cutting plane; and 2) given the cutting planes, find the optimal legislator coordinates. Sections 3 and 4 show solutions for these two subproblems.

3. Finding the Optimal Cutting Plane

Given the p by s matrix, **X**, of legislator coordinates and the p by 1 vector of votes on the jth roll call, $\underline{\mathbf{t}}$, the problem is to find the plane that divides the legislators into two groups such that the number of correct classifications is maximized. Figure 1 shows an example in two dimensions.

Figure 1A. Ten Point Example
Original Positions in Two Dimensions

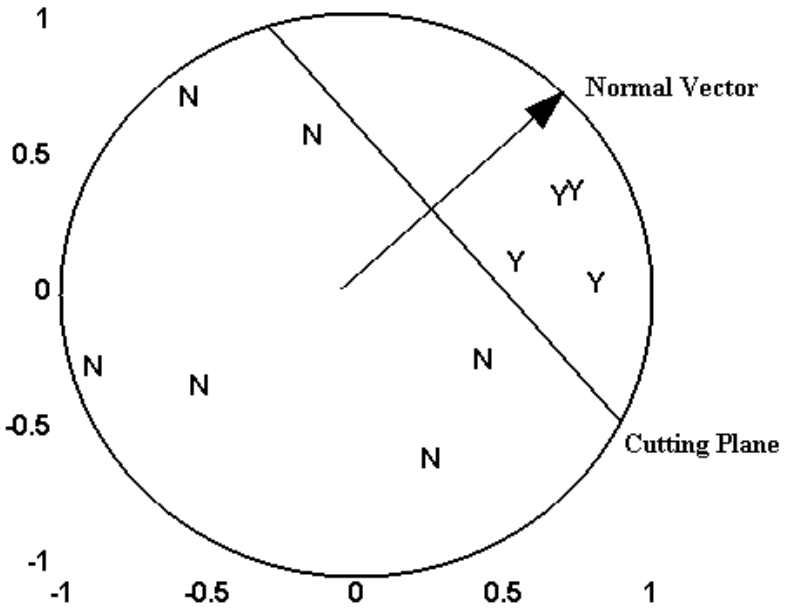


Figure 1B. Ten Point Example
Points Projected Onto Line

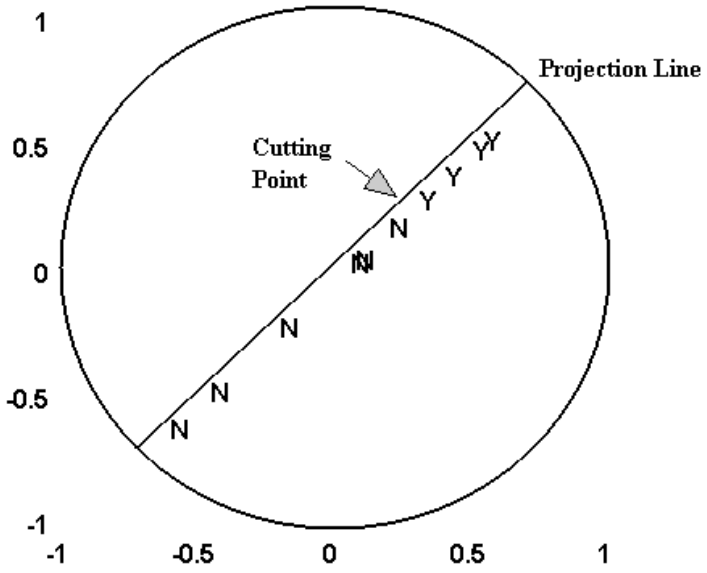


Figure 1 illustrates the fact that the cutting plane problem is equivalent to finding a vector – in this case, $\underline{\mathbf{n}}$ – such that when the legislator points are projected onto the vector a cutting point can be found that maximizes the correct classifications. By definition, all points in the cutting plane are projected onto this cutting point. The problem has two distinct parts which I will discuss in turn. First, given an estimate of the normal vector, I show how to find the optimal plane perpendicular to the normal vector and its associated correct classifications; and second, given an estimated cutting plane, I show how to change the orientation of the plane through the s-dimensional space in order to find a better estimate of the normal vector.

3. a. Calculating the Correct Classifications

Without loss of generality let the legislator coordinates lie within the s dimensional unit hypersphere and let the origin of the space be placed at the centroid of the legislator coordinates; that is, let

$$\sum_{k=1}^{s} x_{ik}^2 \leq 1 \quad , i=1,...,p \qquad \text{and} \qquad \sum_{i=1}^{p} x_{ik} = 0 \quad , k=1,...,s$$

In addition, let $\underline{\mathbf{n}}_j$ be the normal vector for the jth roll call that maximizes correct classifications. Without loss of generality $\underline{\mathbf{n}}_j$ can be constrained to be of unit length; i. e., $\underline{\mathbf{n}}_j'\underline{\mathbf{n}}_j = 1$. The projections (see Figure 1B) are, therefore:

$$\mathbf{X}\underline{\mathbf{n}}_j = \underline{\mathbf{w}} \qquad (1)$$

Note that the elements in the p-length vector, $\underline{\mathbf{w}}$, range from -1 to +1. (To see this, let one of the legislator points be $-\underline{\mathbf{n}}_j$ and another legislator point be $+\underline{\mathbf{n}}_j$.) The elements in $\underline{\mathbf{w}}$ all lie on a line that passes through the origin of the s-dimensional unit hypersphere in the direction of the

normal vector with exit points $-\underline{\mathbf{n}}_j$ and $+\underline{\mathbf{n}}_j$ respectively. Hereafter, I will refer to this line as the projection line.

Let $\underline{\mathbf{n}}_j^*$ be an estimate of $\underline{\mathbf{n}}_j$ and let $\underline{\mathbf{w}}^*$ be the corresponding estimate of $\underline{\mathbf{w}}$. The correct classifications associated with $\underline{\mathbf{n}}_j^*$ can be calculated quite easily. Figure 2 illustrates the method.

Figure 2. Calculating Correct Classification

Actual Voting Pattern

Y Y Y Y Y Y . . . . Y Y * N Y * N Y * N N . . .N N

-1.0 w_1 w_2 w_3 w_4 . . . . 0.0 . . . . w_{p-1} w_p +1.0

Perfect Voting Patterns

$(-1 , w_1)$	**produces**	**nnnnnnnnn.....nn**	**or**	**yyyyyyyyy.....yy**
(w_1 , w_2)	**produces**	**ynnnnnnnn.....nn**	**or**	**nyyyyyyyy.....yy**
(w_2 , w_3)	**produces**	**yynnnnnnn.....nn**	**or**	**nnyyyyyyy.....yy**
(w_3 , w_4)	**produces**	**yyynnnnnn.....nn**	**or**	**nnnyyyyyy.....yy**

etc.

(w_{p-1} , w_p)	**produces**	**yyyyyyyyy.....yn**	**or**	**nnnnnnnnn.....ny**
$(w_p , +1)$	**produces**	**yyyyyyyyy.....yy**	**or**	**nnnnnnnnn.....nn**

For ease of exposition, let the projected legislator coordinates from left to right be denoted in order as w_1 to w_p such that $-1 \le w_1 \le w_2 \le w_3 \le ... \le w_p \le +1$ and the "y"s and "n"s above the dimension line in the Figure indicate how the corresponding legislators voted on the jth roll call. There are p+1 possible regions that the cutting point could be in -- $(-1, w_1)$, (w_1 , w_2), ... , $(w_p , +1)$ -- and for each region there are exactly 2 possible perfect voting patterns for an

overall total of 2(p+1) possible perfect voting patterns as shown in the Figure. However, region (w_p , +1) is redundant since it produces the same perfect patterns as the region (-1 , w_1) so it may be discarded leaving 2p unique perfect voting patterns to consider.

Since there are only 2p perfect patterns, it is a simple matter to compare each perfect pattern with the actual pattern of votes, $\underline{\mathbf{t}}_j$. This can be done very efficiently by first assuming that the cutting point is in the region (-1 , w_1) and calculating the corresponding number of correct classifications. Next assume that the cutting point is in the region (w_1 , w_2). Only one calculation has to be made to get the correct classifications for this cutting point since the only change is that the cutting point has been moved from the left of w_2 to the right of w_2. If there is no missing data, either the correct classification increases by 1 or decreases by 1 when the cutting point is moved from the left of w_2 to the right of w_2. Similar reasoning holds for the remaining points. For each possible cutting point the correct classification corresponding to the two possible perfect patterns can be calculated. The estimated cutting point is set equal to the midpoint of the region for which correct classification is a maximum. For the example shown in Figure 2, placing the cutting point at the position of either of the three asterisks would produce only two classification errors for a correct classification of p-2.

Note that this process is equivalent to moving the cutting plane through the unit hypersphere along the estimated normal vector, $\underline{\mathbf{n}}_j^*$.

3. b. Calculating the Optimal $\underline{\mathbf{n}}_j^*$

Let c* denote the cutting point that maximizes correct classification on the projection line formed by the elements of $\mathbf{X}\underline{\mathbf{n}}_j^* = \underline{\mathbf{w}}^*$. The point c* is therefore:

$$\underline{\mathbf{z}}'\underline{\mathbf{n}}_j^* = \underline{\mathbf{v}}'\underline{\mathbf{n}}_j^* = c^* \qquad (2)$$

Given $\underline{\mathbf{n}_j}^*$ and c*, the estimated cutting plane consists of all points $\underline{\mathbf{v}}$ satisfying equation (2). In order to get a new estimate of $\underline{\mathbf{n}_j}$, the estimated cutting plane given by equation (2) must be rotated through the space in a direction that increases correct classification. I accomplish this by rotating the cutting plane towards the legislator points which are classification errors.

To do this, I create the cutting plane by projecting all the correctly classified legislator points onto the surface of the cutting plane while leaving the incorrectly classified legislators at their original positions. In two dimensions this produces a line through the space made up of correctly classified legislators around which is a scattering of points corresponding to the incorrectly classified legislators (see Figure 3).[1] Much in the spirit of the classic ordinary least squares regression problem, a new cutting plane can then be estimated by simply finding the plane that best fits this set of points using the principle of least squares.

Figure 3A. Cutting Plane Example

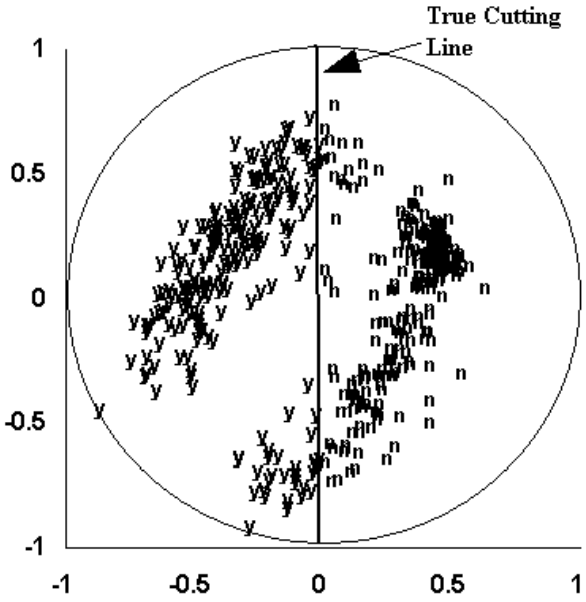


Figure 3B. Cutting Plane Example
Initial Estimate of Cutting Plane

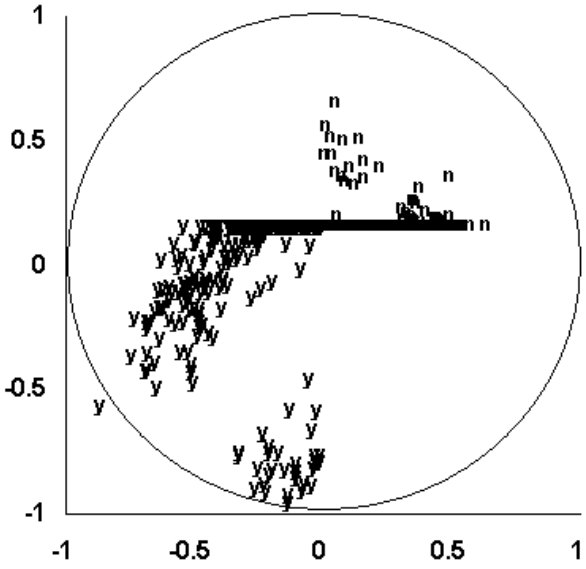


Figure 3C. Cutting Plane Example
2nd Estimate of Cutting Plane

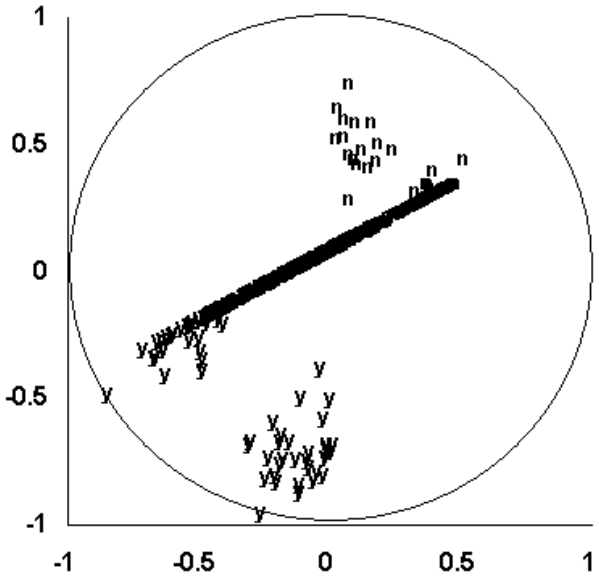


Figure 3D. Cutting Plane Example
10th Estimate of Cutting Plane

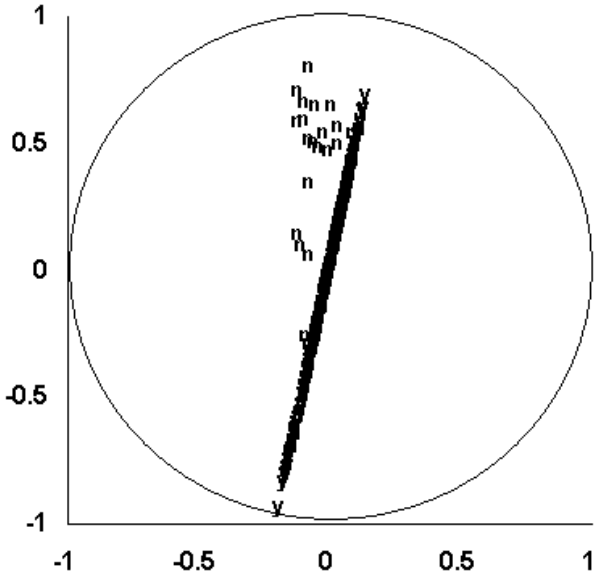
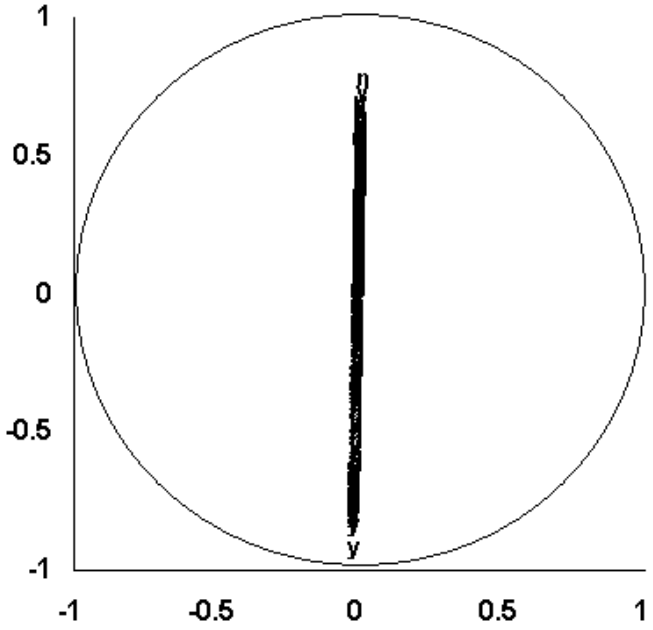


Figure 3E. Cutting Plane Example
35th Estimate of Cutting Plane



To see how this is done, let $\underline{\mathbf{x}}_i$ be the s by 1 vector denoting the ith legislator's point in the space and let w_i be the corresponding point on the projection line from equation (1). Construct a p by s matrix, **V**, as follows: if legislator i is correctly classified, then his point is projected onto the cutting plane and that point becomes the ith row of **V**; if legislator i is incorrectly classified, then his point remains at its original position and that point becomes the ith row of **V**. That is:

$$\begin{aligned} \underline{\mathbf{v}}_i &= \underline{\mathbf{x}}_i + (c^* - w_i)\underline{\mathbf{n}}_j^* \quad \text{if correctly classified} \\ \underline{\mathbf{v}}_i &= \underline{\mathbf{x}}_i \qquad\qquad\qquad\qquad \text{if incorrectly classified} \end{aligned} \tag{3}$$

In the correctly classified case, to see that $\underline{\mathbf{v}}_i$ is on the plane defined in equation (2), note that

$$\underline{\mathbf{n}}_j^{*\prime}\underline{\mathbf{v}}_i = \underline{\mathbf{n}}_j^{*\prime}\underline{\mathbf{x}}_i + \underline{\mathbf{n}}_j^{*\prime}\underline{\mathbf{n}}_j^* (c^* - w_i) = w_i + (c^* - w_i) = c^*$$

because $\underline{\mathbf{n}}_j^{*\prime}\underline{\mathbf{n}}_j^* = 1$ and $(c^* - w_i)$ is a scaler.

Without loss of generality, the centroid of **V** can be placed at the origin. That is, let $\underline{\mathbf{m}}$ be the s length vector of the means of **V**, and let J_p be a p by 1 vector of ones. Define **V*** as

$$\mathbf{V^*} = \mathbf{V} - J_p\underline{\mathbf{m}}' \qquad (4)$$

Panel A of Figure 3 shows a vote in two dimensions which would be perfectly classified by the indicated cutting line. Panel B shows the **V*** produced by using an initial estimate of $\underline{\mathbf{n}}_j^{*\prime}$ = (0 , 1) -- that is, an estimated normal vector perpendicular to the true normal vector. All the "y" and "n" tokens off the plane are classification errors. Clearly, if the plane were rotated counter-clockwise towards the errors a better fit would be obtained. This can be accomplished by fitting a line through the scatterplot of "y"s and "n"s in panel B.

This is accomplished by using a famous result due to Eckart and Young (1936). Eckart and Young addressed the following problem. Let **A** be a p by s matrix of rank s which, by definition, is a set of p points in an Euclidean space of s dimensions. Estimate the best r-dimensional hyperplane -- where r < s -- through this set of points. That is, find a p by s matrix **B** of rank r < s, such that

$$\sum_{i=1}^{p}\sum_{k=1}^{s}(a_{ik} - b_{ik})^2$$

is minimized. Eckart and Young proved that **B** is found by performing a singular value decomposition of **A** , inserting zeroes in place of the s-r smallest singular values, and remultiplying.[2]

To show how this is accomplished with respect to **V***, let the singular value decomposition of **V*** be

$$\mathbf{V^*} = \mathbf{U\Lambda\Theta}' \qquad (5)$$

where **U** is a p by s orthogonal matrix consisting of the first s eigenvectors of the p by p matrix **V*V*′**, **Θ** is an s by s orthogonal matrix consisting of the s eigenvectors of the s by s matrix **V*′V***, and **Λ** is an s by s diagonal matrix containing the singular values in descending order on the diagonal (**V*V*′** and **V*′V*** have the same non-zero eigenvalues – the singular values are the square roots of these eigenvalues). Let $\mathbf{I_s}$ be the s by s identity matrix. By definition, $\mathbf{U'U} = \mathbf{\Theta'\Theta} = \mathbf{I_s}$.[3]

By the Eckart-Young result, the best fitting line through the scatterplot shown in panel B of Figure 3 is found by inserting a zero in place of the second singular value in **Λ** and remultiplying. That is, let $\mathbf{\Lambda}^{\#}$ be the s by s diagonal matrix identical to **Λ** except for the replacement of the sth singular value (by construction, the smallest singular value) by zero, then the estimated hyperplane is:

$$\mathbf{V}^{\#} = \mathbf{U\Lambda}^{\#}\mathbf{\Theta'} \qquad (6)$$

where $\mathbf{V}^{\#}$ will be of rank s-1 by construction.

Let $\underline{\mathbf{n}}_\mathbf{j}^{\#}$ be the normal vector of the hyperplane defined by $\mathbf{V}^{\#}$ and let $\underline{\theta}_\mathbf{s}$ be the sth singular vector (eigenvector) of **Θ** . I will now prove that $\underline{\mathbf{n}}_\mathbf{j}^{\#} = \underline{\theta}_\mathbf{s}$.

By the definition of a plane:

$$\mathbf{V}^{\#}\underline{\mathbf{n}}_\mathbf{j}^{\#} = J_p c^{\#} \qquad (7)$$

where J_p is a p by 1 vector of ones and $c^{\#}$ is a constant. Recall from equation (4) that

$$\sum_{i=1}^{p} v_{ik}^{*} = 0 \quad , k=1,...,s$$

and therefore the p length eigenvectors of **U** in equation (5) must also sum to zero; that is;

$$\sum_{i=1}^{p} u_{ik} = 0 \quad , k=1,...,s$$

From which it follows that the columns of $\mathbf{V}^{\#}$ must also sum to zero:

$$\sum_{i=1}^{p} v_{ik}^{\#} = 0 \quad , k=1,...,s$$

Hence, by simply adding up all p elements of the vectors on either side of the equality in equation (7) it must be the case that $c^{\#} = 0$. Therefore, equation (7) can be rewritten as

$$\mathbf{U}\Lambda^{\#}\Theta'\underline{\mathbf{n}}_j^{\#} = \underline{\mathbf{0}}_p \qquad (8)$$

where $\underline{\mathbf{0}}_p$ is a p length vector of zeroes. Let $\Lambda^{\#-1}$ be an s by s diagonal matrix with diagonal entries that are the reciprocals of the non-zero diagonal entries of $\Lambda^{\#}$. Multiplying both sides of equation (8) by $\Lambda^{\#-1}\mathbf{U}'$

$$\Lambda^{\#-1}\mathbf{U}'\mathbf{U}\Lambda^{\#}\Theta'\underline{\mathbf{n}}_j^{\#} = \Lambda^{\#-1}\mathbf{U}'\underline{\mathbf{0}}_p$$

this reduces to

$$\Theta^{*\prime}\underline{\mathbf{n}}_j^{\#} = \underline{\mathbf{0}}_s$$

where $\underline{\mathbf{0}}_s$ is an s length vector of zeroes, and Θ^* is identical to Θ except the sth column of Θ^* is all zeroes (hence, the sth row of $\Theta^{*\prime}$ is all zeroes). Now, $\underline{\mathbf{n}}_j^{\#}$ cannot be a vector of zeroes since, by definition, $\underline{\mathbf{n}}_j^{\#\prime}\,\underline{\mathbf{n}}_j^{\#} = 1$. Hence, $\underline{\mathbf{n}}_j^{\#} = \underline{\theta}_s$ is a solution for equation (8).

In sum, calculating the optimal $\underline{\mathbf{n}}_j$ consists of the following steps:

1) Obtain a starting estimate of $\underline{\mathbf{n}}_j^*$ using ordinary least squares.
2) Calculate the correct classifications associated with $\underline{\mathbf{n}}_j^*$.
3) Construct $\mathbf{V}^*$ using equations (3) and (4).
4) Perform singular value decomposition of $\mathbf{V}^*$, $\mathbf{U}\Lambda\Theta'$.
5) Use the sth singular vector of Θ , $\underline{\theta}_s$, as the new estimate of $\underline{\mathbf{n}}_j$.
6) Go to (2).

In a perfect case like that shown in Figure 3, this cutting plane procedure will almost always iterate into the true cutting plane. Panels D and E show the process after the 10^{th} and 35^{th} iterations through steps (2) - (5) above.

I say "almost always" because, with perfect data, the rate of convergence is a function of the number of errors. As the number of errors decreases, the mass of the correctly classified choices increases thereby producing very small changes in the newly estimated normal vectors. This can be seen by comparing panels C and D with panels D and E. Indeed, given the right sort of configuration, the convergence can become so slow that it literally gets "lost" in the precision of the computer. Consequently, I have experimented with a number of simple fixes – for example, local grid searches as well as weighting the error so as to speed convergence. However, the basic procedure is so robust that I will limit my Monte-Carlo reports below to it alone so that the reader will have a clear idea of how well it works "barefoot" without any enhancements.

Table 1 shows a Monte-Carlo study of the cutting plane procedure using perfect data for 100 legislators and 500 roll calls for 2 through 10 dimensions. Results for one dimension are not shown since correct classification will always be 100% if perfect data is used. The 100 legislators and 500 normal vectors were randomly drawn from a uniform distribution through the unit hypersphere. The cutting points along the projection line were also randomly drawn but in such a way so as to produce an average majority margin of about 62 percent (typical of congressional roll call data).[4] A maximum of 100 iterations through steps (2) - (5) above were allowed.

Table 1

Monte-Carlo Tests of Cutting Plane Procedure
100 Legislators and 500 Votes

Number of Dimensions	Trials	Average Majority Margin[a]	Average Number of Errors[b]	Average Percent Correctly Classified[c]
2	10	61.1	8.9	99.98
3	10	62.2	16.8	99.97
4	10	61.9	12.5	99.98
5	10	62.2	16.7	99.97
6	10	63.1	18.9	99.96
7	10	62.5	17.7	99.96
8	10	62.8	15.1	99.97
9	10	62.8	14.0	99.97
10	10	62.8	13.2	99.97

[a] Average margin across all 10 trials or 5000 total votes.

[b] Average across all 10 trials.

[c] Average across all 10 trials or 500,000 total choices.

The cutting plane procedure performs very well. The number of dimensions does not appear to play any role in the accuracy of the procedure. For example, for the ten trials in 10 dimensions, the 5000 total estimated $\underline{\mathbf{n}}_j^*$'s correctly classified 499,868 of 500,000 choices (99.97 percent). As I noted above, this accuracy rate is a baseline; it can be further improved with local searches and other enhancements.

When error is present the cutting plane procedure converges very quickly. In this context I am using "error" in a very artificial sense because I have not stated any behavioral model of the legislators. It simply means that, given a legislator configuration and an observed pattern of yea's and nay's, find the cutting plane that maximizes correct classification. The pattern might be produced by a probabilistic model of legislator voting or it might be a lower space projection of perfect voting in a higher dimensional space. An example is shown in Figure 4 which uses the same configuration of legislator ideal points as Figure 3. The choices of 78 of the 435 legislators have been modified so that they are "errors" – "N's" on the "Y" side of the true cutting line and "Y's" on the "N" side of the true cutting line. The cutting plane procedure converges on the 30^{th} iteration as shown in Panel D. As shown by Panels B and C, in the error case the converged cutting plane may not be the one that maximizes classification – however, it will invariably be very close to the optimal cutting plane. This is easily dealt with by simply storing the iteration record and using the normal vector corresponding to the best classification. This works very well in practice.

Figure 4A. Error Example
78 Errors With True Cutting Line

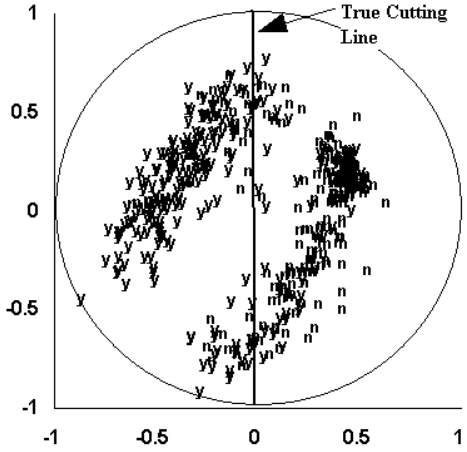


Figure 4B. Error Example
73 Errors at 10th Estimate (Minimum)

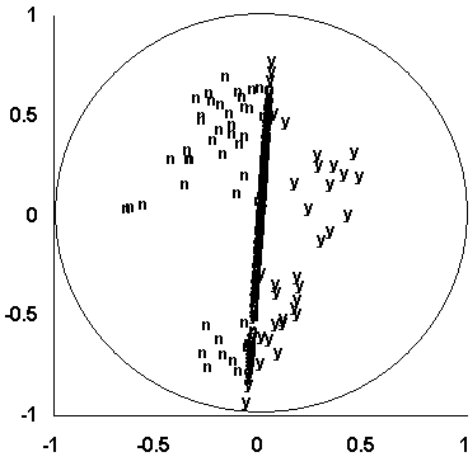


Figure 4C. Error Example
73 Errors at 20th Estimate (Minimum)

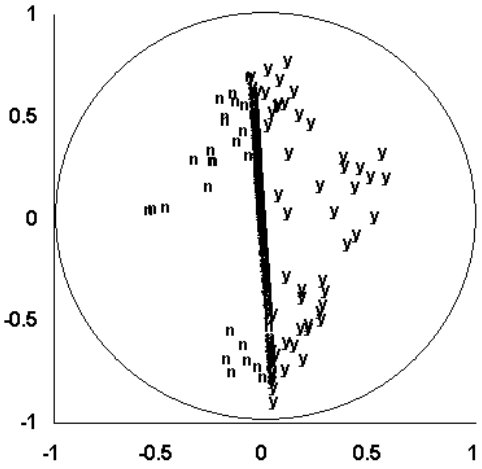
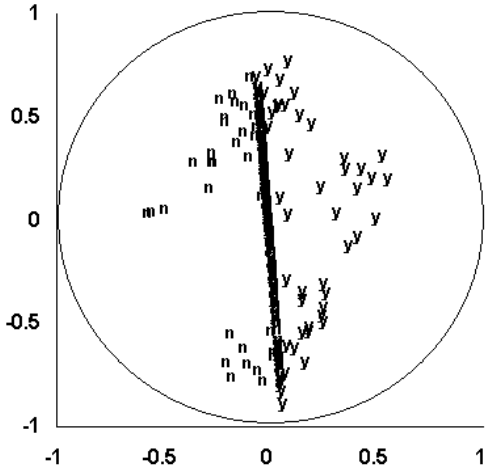


Figure 4D. Error Example
75 Errors at 30th Estimate (Converged)



3.c. The Relationship of the Cutting Plane Procedure to Probit Analysis

Given a simple two category dependent variable and a set of fixed independent variables, the cutting plane procedure can be used to estimate a vector of coefficients for the independent variables that maximizes correct classification of the dependent variable. In this instance, with the independent variables scaled so as to lie within a unit hypersphere, the normal vector, $\underline{\mathbf{n}}_j^*$,

produced by the cutting plane procedure, plays the role of the coefficient vector, $\underline{\beta}$, in a standard probit, logit, or linear probability analysis.

In order to study the relationship between the cutting plane procedure and a standard probit analysis, I performed a Monte-Carlo study in which I created artificial data that fit the model:

$$\underline{\mathbf{y}}^* = \mathbf{X}\underline{\beta} + \underline{\varepsilon}$$

where **X** is a p by s matrix of randomly generated independent variables scaled so as to be within a unit hypersphere;[5] $\underline{\beta}$ is a randomly generated s length vector of coefficients with $\underline{\beta}'\underline{\beta} = 1$;[6] $\underline{\varepsilon}$ is a p length vector of error terms; and $\underline{\mathbf{y}}^*$ is the unobserved latent dependent variable. The categorical dependent variable, $\underline{\mathbf{y}}$, was created by picking a cutpoint on $\underline{\mathbf{y}}^*$ and defining all observations above the cutpoint as one category and all observations below the cutpoint as the second category. I then analyzed **X** and $\underline{\mathbf{y}}$ with probit and the cutting plane procedure. A portion of the study is shown in Table 2.

Table 2

Monte-Carlo Comparison of Cutting Plane Procedure and Probit Analysis

Normally Distributed Error (σ=.2): p=100

Number Variables (s)	True Margin	Correlation Probit With True	Correlation Cutting Plane With True	Correlation Cutting Plane With Probit	Probit Percent Correctly Classified	Cutting Plane Correctly Classified
10	50-50	.946	.949	.995	85	87.9
10	20-80	.924	.913	.989	88.3	90.3
5	50-50	.985	.979	.997	85.3	87.8
Uniformly Distributed Error: p=100						
10	50-50	.824	.825	.996	72.3	75.4
10	20-80	.564	.579	.991	72.4	75.0
5	50-50	.917	.912	.997	72.8	75.8

Asymmetric Error: p=100 and s=10

True Margin	Type of Error	Correlation Probit With True	Correlation Cutting Plane With True	Correlation Cutting Plane With Probit	Probit Percent Correctly Classified	Cutting Plane Correctly Classified
50-50	Chi-Square	.913	.920	.992	86.6	89.7
50-50	Bi-Modal	.778	.796	.997	75.2	78.4
20-80	Error Near End	.779	.863	.983	85.2	90.0

Table 2 shows three sets of experiments using normal, uniform, and asymmetric error, respectively. All entries are the average of 10 trials. In the first portion of the Table the number of independent variables was set equal to either 5 or 10 (s=5 or s=10) with 100 observations (p=100). The randomly drawn $\underline{\beta}$ -- the normal vector – was used to create the true

latent dimension and a cutpoint was selected so that the true margin was 50-50. Normal random error was then drawn and added to $\mathbf{X}\underline{\beta}$ to get the noisy latent dimension and the categories were adjusted vis a vis the true cutpoint. The column "True Margin" is the margin before the addition of the error. The next three columns of the Table report the Pearson correlations between: the true $\underline{\beta}$ and the vector of coefficients from the Probit analysis; the true $\underline{\beta}$ and the estimated normal vector, $\underline{\mathbf{n}_j}^*$, from the cutting plane procedure; and the Probit coefficients and $\underline{\mathbf{n}_j}^*$. Finally, the last two columns show the percent correct classifications of Probit and the cutting plane procedure, respectively.[7]

With normally distributed and uniformly distributed error the non-parametric cutting plane procedure recovers essentially the same set of coefficients as Probit. This seems sensible in that as long as the underlying error distribution is symmetric so that the frequency of error diminishes with distance from the cutting plane, then the cutting plane procedure should produce results very similar to those of a parametric procedure like Probit.

The last set of results shown in Table 2 are those for three asymmetric error distributions: chi-square with one degree of freedom; a bimodal distribution where the frequency of error peaks midway between the cutpoint and the ends of the dimension; and one in which the error is clustered near only one end of the dimension. Even though the two procedures are recovering similar vectors, when the error distribution is not normal and not symmetric, the cutting plane procedure comes closer to recovering the true vector of coefficients than does Probit.

Table 3 shows an empirical comparison of the cutting plane procedure and Probit analysis. The sample is 231 Republican members of the House of Representatives[8] and the dependent variable is whether or not they signed up as co-sponsors of a minimum wage increase.[9] The independent variables measure the liberalness of the representative (the

NOMINATE scores; see Poole and Rosenthal, 1991, 1997, for details) and some characteristics of representative's congressional district (percent rural, percent Black, and median family income). (The independent variables were put in standard deviation form to facilitate comparisons.) The standardized Probit coefficients and the cutting plane coefficients are very close – the simple Pearson correlation is .961. Substantively, the coefficients in Table 3 indicate that Republican moderates from poorer, urban districts support raising the minimum wage.

Table 3

Empirical Comparison of Probit and Cutting Plane Procedure
22 Republican Defectors on Minimum Wage: April 1996

Dependent Variable = 1 if Support Raising Minimum Wage; 0 if Oppose
Independent Variables Expressed in Standard Deviation Form
Margin 22 - 209

Variable	Probit Coefficient	Standardized Probit Coefficients[d]	T-Value	Cutting Plane Coefficients	Boot-Strapped T-Values[e]
Constant	1.722	---	9.253	---	---
NOMINATE[a] 1st Dimension	12.211	.911	3.508	.876	3.681
NOMINATE[a] 2nd Dimension	0.785	.059	0.266	.266	.767
Rural[b]	4.157	.310	1.922	.338	2.113
Black[b]	0.337	.025	0.175	.051	.338
Median Income[c]	3.568	.266	1.436	.212	.964

Probit Log Likelihood = -54.101
Percent Correctly Classified By Probit = 90.9 (210 of 231)
Percent Correctly Classified By Cutting Plane Procedure = 91.8 (212 of 231)
Correlation Between Cutting Plane Coefficients and Standardized Probit Coefficients = .961

---

[a] Unadjusted NOMINATE scores range from -1.0 to +1.0
[b] Unadjusted data expressed as a percentage
[c] Unadjusted data expressed in dollars
[d] The sum of the squared coefficients equals 1
[e] Based upon 100 trials

In order to obtain standard errors for the cutting plane procedure, I performed a simple bootstrap analysis.[10] The standard errors were then used to obtain the reported "t-values". Note that the pattern of significance for the non-parametric cutting plane coefficients is the same as that for the Probit coefficients. Monte-Carlo work with artificial data suggests that the cutting plane coefficients will have nearly identical patterns of significance (using bootstrapping) with those produced by a Probit analysis when the underlying error distribution is symmetric.

3.d. Conjecture on the Relationship of the Cutting Plane Procedure to Ordered Probit and Multi-Choice Analysis

I conjecture that the cutting plane procedure is easily generalized to the case of ordered Probit. The only modification necessary is to change the way correct classifications are counted. For example, suppose there are four categories – y, n, a, b – and the order is y > n > a > b (or its mirror image). Given an estimate of the normal vector, $\underline{\mathbf{n}}_j^*$, the problem is to find three parallel cutting planes that divide the space into four regions so as to maximize the correct classifications. This can be solved using a modification of the classification procedure shown in Section 3.a.

To illustrate, consider one of the intermediate categories, "n". Given the projection line, equation (1), the problem now is to find <u>two</u> cutpoints rather than one. Lump the other three categories – y, a, b – into a <u>non</u>-"n" category. Denote the non-"n" category as "c". Using the same logic discussed is Section 3.a, search all the patterns of the form:

cnnnn…nnccccccccc
ccnnn…nnccccccccc
cccnn…nnccccccccc
etc.
ccccc…ccccnnnccc

etc.

Given the restriction that there must be at least one "c" to the left of the left-most "n" and at least two "c"s to the right of the right-most "n", and at least one "n", then there are exactly $\binom{p-2}{2}$ possible pairs of cutpoints which define $\binom{p-2}{2}$ possible patterns of "c"s and "n"s. Given these two cutpoints, the simple one point procedure can be used to find the last cutpoint.

Because the three cutting planes are parallel and have the same normal vector, the correctly classified observations can be projected onto a "common" plane which is placed through the origin of the space rather than through the cutpoint, c^*, as shown in equation (3). The errors are then translated so that they are the same distance and orientation from the "common" plane as the plane for which they are an error.

In a multiple choice context where there is no natural ordering of the choices, the parametric approach consists of estimating separate probabilities for each of the choices utilizing a common set of independent variables. For example, in the 1980 U. S. presidential election there were four choices – Reagan, Carter, Anderson, and not-voting. The probability that an eligible voter voted for Reagan is the product of the conditional probability that, given the person decides to vote, she votes for Reagan times the probability that she votes. Using a logit model, three vectors of coefficients will be estimated – one for the vote/not-vote decision, and two for the candidate choices (only two are needed since probabilities add to one).[11]

Because the probability of making a choice rises in the direction of the coefficient vector, in terms of the cutting plane procedure, this is equivalent to estimating <u>three</u> normal vectors – one for the binary choice vote/not-vote; and two for the voters, one for the binary choice Reagan/not-Reagan; and one for the binary choice Anderson/not-Anderson (using Carter as the

omitted choice). Geometrically this approach is possible because, from some interior point outward along a vector from the origin, there should be only one choice. However, the cutting plane procedure is linear – that is, it uses straight cuts to divide up the hypersphere when curved boundaries between the choice regions may be more appropriate because the hypersphere must be divided into mutually exclusive regions. Consequently, the cutting plane procedure should not do as well in this context as it should with ordered Probit. Nevertheless, it should provide a very useful benchmark for comparison with parametric methods.

4. Finding the Optimal Legislator Coordinates

Given the q by s matrix, **N**, of normal vectors and the q by 1 vector of votes of the ith legislator, $\underline{\mathbf{t}}_i$, the problem is to find the legislator point, $\underline{\mathbf{x}}_i$, which minimizes the classification error. Figure 5 shows an example in two dimensions.

Figure 5. Legislator Example: nnnyn

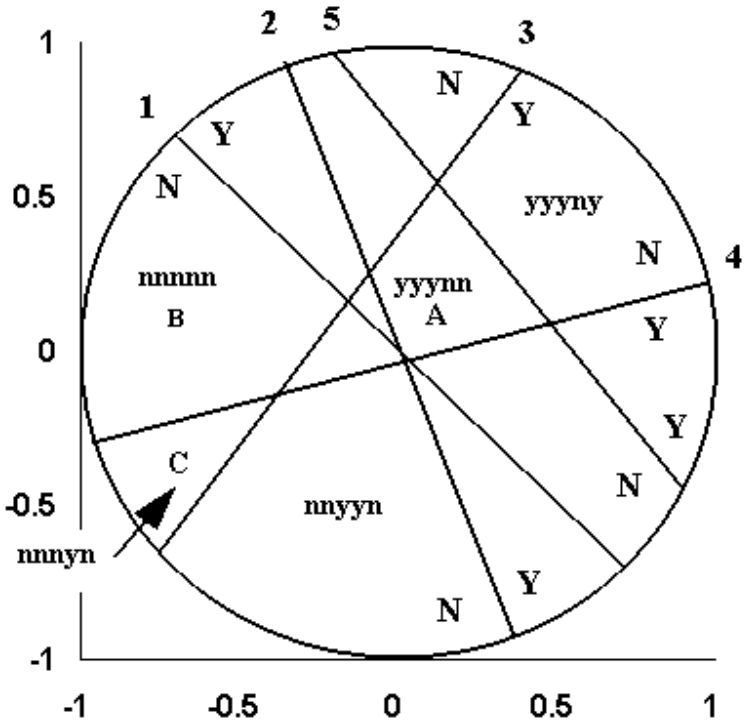


Figure 5 shows five cutting lines indicated by the numbering at the rim of the circle. The "Y" and "N" on either side of each cutting line indicates how a legislator on that side of the cutting line should vote – "yea" or "nay" respectively. The five cutting lines divide up the space into 13 regions and each of these 13 regions can be characterized by a unique vector of votes. The voting patterns for several regions are shown in the figure. For example, the region near the center of the circle containing the point "A" corresponds to a voting pattern of yyynn.

Given a legislator's pattern of votes, in this case nnnyn (technically, $\underline{\mathbf{t}}_i' = [\text{nnnyn}]$), the problem is to find the region in Figure 5 that maximizes the correct classification. In this example the point "C" is located in the region corresponding to perfect classification. Suppose the initial estimate of the legislator's coordinates is at the origin, point "A" in the Figure. This

initial estimate is very poor as it only correctly classifies one of the five votes. The problem is to move the point representing the legislator in a direction that increases the number of correct classifications.

Below I show a method for finding the maximum classification point along any arbitrary line passing through the space. This method is used to move the legislator point through the space in a city-block fashion by searching along a line parallel to the first dimension and then solving for the point along this line which maximizes classification. Then the legislator point is moved along a line through this new point but parallel to the second dimension. This is done for each dimension in turn and can be repeated as many times as desired. This always converges to a point for which the coordinates are at a local maximum in terms of classification. That is, the point cannot be moved parallel to any dimension and have the correct classifications increase.

Let $\underline{\mathbf{x}}_i^{(h)}$ be the initial estimate for legislator i where "h" is the iteration number (1, 2, 3, etc.) and let $\underline{\mathbf{x}}_i^{(a)}$ be a second point. The problem is to find a new estimate, $\underline{\mathbf{x}}_i^{(h+1)}$, on the line passing through $\underline{\mathbf{x}}_i^{(h)}$ and $\underline{\mathbf{x}}_i^{(a)}$ which increases correct classification. Using equation (1), the projection of $\underline{\mathbf{x}}_i^{(h)}$ onto the jth normal vector is:

$$\underline{\mathbf{x}}_i^{(h)\prime}\underline{\mathbf{n}}_j = w_{ij}^{(h)} \qquad (9)$$

Similarly, the projection of the second point onto the jth normal vector is $w_{ij}^{(a)}$. These projections correspond to a correct classification on roll call j depending upon which side of the cutpoint, c_j, they fall. There are six possible orderings of $w_{ij}^{(h)}$, $w_{ij}^{(a)}$, and c_j. For each ordering there are two possible classification outcomes for a total of 12 cases. Table 4 shows each case.

Table 4

Case	Ordering	Classification h	Classification a	Limits of α That Correctly Project $x_i^{(h+1)}$
1.	$-1 < c_j < w_{ij}^{(h)} < w_{ij}^{(a)} < +1$	C[1]	C	$\frac{c_j - w_{ij}^{(h)}}{w_{ij}^{(a)} - w_{ij}^{(h)}} < \alpha_j < \frac{1 - w_{ij}^{(h)}}{w_{ij}^{(a)} - w_{ij}^{(h)}}$
2.	$-1 < c_j < w_{ij}^{(h)} < w_{ij}^{(a)} < +1$	I	I	$\frac{-1 - w_{ij}^{(h)}}{w_{ij}^{(a)} - w_{ij}^{(h)}} < \alpha_j < \frac{c_j - w_{ij}^{(h)}}{w_{ij}^{(a)} - w_{ij}^{(h)}}$
3.	$-1 < c_j < w_{ij}^{(a)} < w_{ij}^{(h)} < +1$	C	C	$\frac{1 - w_{ij}^{(h)}}{w_{ij}^{(a)} - w_{ij}^{(h)}} < \alpha_j < \frac{c_j - w_{ij}^{(h)}}{w_{ij}^{(a)} - w_{ij}^{(h)}}$
4.	$-1 < c_j < w_{ij}^{(a)} < w_{ij}^{(h)} < +1$	I	I	$\frac{c_j - w_{ij}^{(h)}}{w_{ij}^{(a)} - w_{ij}^{(h)}} < \alpha_j < \frac{-1 - w_{ij}^{(h)}}{w_{ij}^{(a)} - w_{ij}^{(h)}}$
5.	$-1 < w_{ij}^{(h)} < w_{ij}^{(a)} < c_j < +1$	C	C	$\frac{-1 - w_{ij}^{(h)}}{w_{ij}^{(a)} - w_{ij}^{(h)}} < \alpha_j < \frac{c_j - w_{ij}^{(h)}}{w_{ij}^{(a)} - w_{ij}^{(h)}}$
6.	$-1 < w_{ij}^{(h)} < w_{ij}^{(a)} < c_j < +1$	I	I	$\frac{c_j - w_{ij}^{(h)}}{w_{ij}^{(a)} - w_{ij}^{(h)}} < \alpha_j < \frac{1 - w_{ij}^{(h)}}{w_{ij}^{(a)} - w_{ij}^{(h)}}$
7.	$-1 < w_{ij}^{(a)} < w_{ij}^{(h)} < c_j < +1$	C	C	$\frac{c_j - w_{ij}^{(h)}}{w_{ij}^{(a)} - w_{ij}^{(h)}} < \alpha_j < \frac{-1 - w_{ij}^{(h)}}{w_{ij}^{(a)} - w_{ij}^{(h)}}$
8.	$-1 < w_{ij}^{(a)} < w_{ij}^{(h)} < c_j < +1$	I	I	$\frac{1 - w_{ij}^{(h)}}{w_{ij}^{(a)} - w_{ij}^{(h)}} < \alpha_j < \frac{c_j - w_{ij}^{(h)}}{w_{ij}^{(a)} - w_{ij}^{(h)}}$
9.	$-1 < w_{ij}^{(h)} < c_j < w_{ij}^{(a)} < +1$	C	I	$\frac{-1 - w_{ij}^{(h)}}{w_{ij}^{(a)} - w_{ij}^{(h)}} < \alpha_j < \frac{c_j - w_{ij}^{(h)}}{w_{ij}^{(a)} - w_{ij}^{(h)}}$
10.	$-1 < w_{ij}^{(h)} < c_j < w_{ij}^{(a)} < +1$	I	C	$\frac{c_j - w_{ij}^{(h)}}{w_{ij}^{(a)} - w_{ij}^{(h)}} < \alpha_j < \frac{1 - w_{ij}^{(h)}}{w_{ij}^{(a)} - w_{ij}^{(h)}}$
11.	$-1 < w_{ij}^{(a)} < c_j < w_{ij}^{(h)} < +1$	C	I	$\frac{1 - w_{ij}^{(h)}}{w_{ij}^{(a)} - w_{ij}^{(h)}} < \alpha_j < \frac{c_j - w_{ij}^{(h)}}{w_{ij}^{(a)} - w_{ij}^{(h)}}$
12.	$-1 < w_{ij}^{(a)} < c_j < w_{ij}^{(h)} < +1$	I	C	$\frac{c_j - w_{ij}^{(h)}}{w_{ij}^{(a)} - w_{ij}^{(h)}} < \alpha_j < \frac{-1 - w_{ij}^{(h)}}{w_{ij}^{(a)} - w_{ij}^{(h)}}$

[1] "C" is correctly classified; "I" is incorrectly classified.

For example, in case 1 both $\underline{\mathbf{x}}_i^{(h)}$ and $\underline{\mathbf{x}}_i^{(a)}$ project to the right of c_j and are on the correct side of the cutting plane for the jth roll call and are therefore correctly classified. Case 2 is the same geometrically only now $\underline{\mathbf{x}}_i^{(h)}$ and $\underline{\mathbf{x}}_i^{(a)}$ are on the wrong side of the cutting plane and are therefore projected as classification errors. Cases 1 to 8 represent no change in classification from moving the legislator point from $\underline{\mathbf{x}}_i^{(h)}$ to $\underline{\mathbf{x}}_i^{(a)}$. For $\underline{\mathbf{x}}_i^{(a)}$ to be an improvement over $\underline{\mathbf{x}}_i^{(h)}$, the number of cases 10 and 12 must be greater than the number of cases 9 and 11.

Consider the effect of moving $\underline{\mathbf{x}}_i^{(a)}$ further from $\underline{\mathbf{x}}_i^{(h)}$. This has no effect on cases 1, 2, and 7 - 12. Only those cases where $\underline{\mathbf{x}}_i^{(a)}$ is between $\underline{\mathbf{x}}_i^{(h)}$ and c_j – cases 3, 4, 5, and 6 – are affected. Depending upon how far $\underline{\mathbf{x}}_i^{(a)}$ is moved away from $\underline{\mathbf{x}}_i^{(h)}$, case 3 could change to case 11 increasing the error by one, case 5 could change to case 9 also increasing the error by one, case 4 could change to case 12 decreasing the error by one, and case 6 could change to case 10 also decreasing the error by one. A similar analysis of the effect of moving $\underline{\mathbf{x}}_i^{(a)}$ towards $\underline{\mathbf{x}}_i^{(h)}$ can also be done.

More generally, consider the line equation:

$$\underline{\mathbf{x}}_i^{(h+1)} = \underline{\mathbf{x}}_i^{(h)} + \alpha(\underline{\mathbf{x}}_i^{(a)} - \underline{\mathbf{x}}_i^{(h)}) \qquad (10)$$

which, when projected onto the jth normal vector, becomes:

$$w_{ij}^{(h+1)} = w_{ij}^{(h)} + \alpha(w_{ij}^{(a)} - w_{ij}^{(h)}) \qquad (11)$$

For a single roll call, it is easy to solve for α; these are shown in Table 4 for all 12 cases. For example, for case 2, α must be chosen so that the projection of $\underline{\mathbf{x}}_i^{(h+1)}$, $w_{ij}^{(h+1)}$, is in the region $(-1, c_j)$.

Given $\underline{\mathbf{x}}_i^{(h)}$ and $\underline{\mathbf{x}}_i^{(a)}$, Table 4 can be used to find the limits of α for each roll call. Let the upper and lower limits for the jth roll call be U_{ij} and L_{ij} respectively. The correct classification associated with $\underline{\mathbf{x}}_i^{(h)}$ can be obtained by setting $\alpha=0$ and counting the number of roll calls for

which $0 \in (L_{ij}, U_{ij})$. Similarly, the correct classification associated with $\underline{\mathbf{x}}_i^{(a)}$ is obtained by setting $\alpha=1$ and counting the number of roll calls for which $1 \in (L_{ij}, U_{ij})$. In general, define

$$\delta_{ij} = 1 \quad \text{if} \quad \alpha \in (L_{ij}, U_{ij})$$

$$\delta_{ij} = 0 \quad \text{if} \quad \alpha \notin (L_{ij}, U_{ij})$$

and the correct classification is simply

$$\delta(\alpha) = \sum_{j=1}^{q} \delta_{ij} \qquad (12)$$

The α that maximizes $\delta(\alpha)$, the number of correct classifications, can be calculated in a simple manner. First, compute the L_{ij} and U_{ij} for each roll call. Second, rank order the L_{ij} and U_{ij} and use the classification algorithm described in section 3.a above to calculate the optimal α. Here the L_{ij} play the role of "y" and the U_{ij} play the role of "n". For example, if there exists an α that results in perfect classification, the ordering of L's and U's will look like (dropping the i subscript to reduce clutter and numbering left to right for convenience):

$$L_1 < L_2 < L_3 < ... < L_q < U_1 < U_2 < U_3 < ... < U_q$$

that is, all the L_j will be less than all the U_j. In this example, perfect classification, $\delta(\alpha) = q$, results from $\alpha \in (L_q, U_1)$. Table 5 shows a numerical example using the configuration of 5 cutting lines shown in Figure 5.

Table 5

Numerical Example From Figure 5

$$\textbf{Using} \quad \mathbf{N} = \begin{bmatrix} .705 & .710 \\ .938 & .346 \\ .834 & -.551 \\ .272 & -.962 \\ .808 & .589 \end{bmatrix} \quad \textbf{and} \quad \underline{\mathbf{c}} = \begin{bmatrix} -.010 \\ -.032 \\ -.203 \\ .076 \\ .021 \end{bmatrix}$$

First Iteration: Set $\underline{\mathbf{x}}_i^{(1)} = \begin{bmatrix} \mathbf{O} \\ \mathbf{O} \end{bmatrix}$ and $\underline{\mathbf{x}}_i^{(a)} = \begin{bmatrix} 0.01 \\ 0 \end{bmatrix}$

	Lower Limits		**Upper Limits**
L_1	**-141.844**	U_1	**-1.349**
L_2	**-106.610**	U_2	**-3.411**
L_3	**-119.904**	U_3	**-24.322**
L_4	**28.010**	U_4	**367.647**
L_5	**-123.762**	U_5	**2.649**

Rank Order

$L_1 < L_5 < L_3 < L_2 < U_3 < U_2 < U_1 < U_5 < L_4 < U_4$

Correct Classifications

For $\underline{\mathbf{x}}_i^{(1)}$, $\alpha = 0 \in (U_1 , U_5)$ and $\delta(0) = 1$
For $\underline{\mathbf{x}}_i^{(a)}$, $\alpha = 1 \in (U_1 , U_5)$ and $\delta(1) = 1$
For $\underline{\mathbf{x}}_i^{(2)} = \underline{\mathbf{x}}_i^{(1)} + \alpha(\underline{\mathbf{x}}_i^{(a)} - \underline{\mathbf{x}}_i^{(1)})$, $\alpha \in (L_2 , U_3)$ and $\delta(\alpha \in (L_2 , U_3)) = 4$

Compute $\underline{\mathbf{x}}_i^{(2)}$ (Point "B" in Figure 5)

Set $\alpha = (L_2 + U_3)/2 = (-106.610 + -24.322)/2 = -65.466$

$$\underline{\mathbf{x}}_i^{(2)} = \begin{bmatrix} \mathbf{O} \\ \mathbf{O} \end{bmatrix} + (-65.466) \left\{ \begin{bmatrix} 0.01 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\} = \begin{bmatrix} -.655 \\ 0 \end{bmatrix}$$

Table 5 (Cont.)

Second Iteration: $\underline{\mathbf{x}}_i^{(2)} = \begin{bmatrix} -.655 \\ 0 \end{bmatrix}$ and $\underline{\mathbf{x}}_i^{(a)} = \begin{bmatrix} -.655 \\ 0.01 \end{bmatrix}$

Lower Limits		Upper Limits	
L_1	-75.841	U_1	63.666
L_2	-111.541	U_2	168.231
L_3	-62.276	U_3	82.276
L_4	-122.461	U_4	-26.430
L_5	-79.972	U_5	93.441

Rank Order

$L_4 < L_2 < L_5 < L_1 < L_3 < U_4 < U_1 < U_3 < U_5 < U_2$

Correct Classifications

For $\underline{\mathbf{x}}_i^{(2)}$, $\alpha = 0 \in (U_4 , U_1)$ and $\delta(0) = 4$
For $\underline{\mathbf{x}}_i^{(a)}$, $\alpha = 1 \in (U_4 , U_1)$ and $\delta(1) = 4$
For $\underline{\mathbf{x}}_i^{(3)} = \underline{\mathbf{x}}_i^{(2)} + \alpha(\underline{\mathbf{x}}_i^{(a)} - \underline{\mathbf{x}}_i^{(2)})$, $\alpha \in (L_3 , U_4)$ and $\delta(\alpha \in (L_3 , U_4)) = 5$

Compute $\underline{\mathbf{x}}_i^{(3)}$ (Point "C" in Figure 5)

Set $\alpha = (L_3 + U_4)/2 = (-62.276 + -26.230)/2 = -44.353$

$$\text{New } \underline{\mathbf{x}}_i^{(3)} = \begin{bmatrix} -.655 \\ 0 \end{bmatrix} + (-44.353) \left\{ \begin{bmatrix} -.655 \\ 0.01 \end{bmatrix} - \begin{bmatrix} -.655 \\ 0 \end{bmatrix} \right\} = \begin{bmatrix} -.655 \\ -.444 \end{bmatrix}$$

The starting estimate (h=1) $\underline{\mathbf{x}}_i^{(1)}$, is placed at the origin – point "A" in Figure 5 – and the second point, $\underline{\mathbf{x}}_i^{(a)}$, is placed just to the right of $\underline{\mathbf{x}}_i^{(1)}$. The upper and lower limits (computed from Table 4) along with their rank order are shown in the Table. The rank ordering is almost a perfect pattern in that 4 of the lower limits are below the 5 upper limits; only L_4 is wrongly placed producing one classification error. Consequently, the point resulting from using $\alpha \in (L_2 ,$

U_3) , $\underline{\mathbf{x}}_i^{(2)}$, point "B" in Figure 5, only has one classification error with 4 correct classifications. (In practice, α is set equal to the midpoint; in this case,($L_2 + U_3$)/2 .) Note that in Figure 5 point "B" is on the wrong side of the cutting line for roll call 4 in the region associated with the pattern nnnnn.

For the second iteration, h=2, the starting estimate is $\underline{\mathbf{x}}_i^{(2)}$ and the second point, $\underline{\mathbf{x}}_i^{(a)}$, is placed just below $\underline{\mathbf{x}}_i^{(2)}$ so that the resulting line is parallel to the second dimension. The upper and lower limits for the second iteration along with their rank order are shown in the Table. The rank ordering is now a perfect pattern with all 5 lower limits below the 5 upper limits so that there are no classification errors. The point resulting from using $\alpha \in (L_3 , U_4)$, $\underline{\mathbf{x}}_i^{(3)}$, point "C" in Figure 5, has 5 correct classifications and no classification error.

The search for the optimal $\underline{\mathbf{x}}_i$ is conducted in a city-block manner. In the first iteration the search is along a line through the origin with all but the first dimension coordinates in $\underline{\mathbf{x}}_i^{(1)}$ and $\underline{\mathbf{x}}_i^{(a)}$ set to zero. In the second iteration, the first dimension coordinates are all set equal to the value corresponding to the optimal first dimension value and the 3^{rd}, 4^{th}, …, s^{th} dimensional coordinates in $\underline{\mathbf{x}}_i^{(2)}$ and $\underline{\mathbf{x}}_i^{(a)}$ are all set equal to zero. The search is along the corresponding line through $\underline{\mathbf{x}}_i^{(2)}$ and $\underline{\mathbf{x}}_i^{(a)}$ which is orthogonal to the first dimension. In the third iteration, the first and second dimension coordinates are set equal to the optimal values from the first and second iterations respectively, and the 4^{th}, 5^{th}, …, s^{th} dimensional coordinates in $\underline{\mathbf{x}}_i^{(3)}$ and $\underline{\mathbf{x}}_i^{(a)}$ are all set equal to zero. The search is along the corresponding line through $\underline{\mathbf{x}}_i^{(3)}$ and $\underline{\mathbf{x}}_i^{(a)}$ which is orthogonal to the second dimension. This process continues in the same fashion through the s^{th} dimension. Since the search for the optimal $\underline{\mathbf{x}}_i$ is being done city-block-wise, dimensions 1 to s can now be searched again.

In sum, calculating the optimal $\underline{\mathbf{x}}_i$ consists of the following steps:

1) Obtain a realistic starting estimate, $\underline{\mathbf{x}}_i^{(1)}$ (or set $\underline{\mathbf{x}}_i^{(1)}$ equal to the origin, that is, $\underline{\mathbf{x}}_i^{(1)} = \mathbf{0}$).

2) Set $\underline{\mathbf{x}}_i^{(a)\prime} = (0.01, x_{i2}^{(1)}, x_{i3}^{(1)}, x_{i4}^{(1)}, x_{i5}^{(1)}, \ldots, x_{is}^{(1)})$, find optimal α and $\underline{\mathbf{x}}_i^{(2)} = \underline{\mathbf{x}}_i^{(1)} + \alpha(\underline{\mathbf{x}}_i^{(a)} - \underline{\mathbf{x}}_i^{(1)})$.

3) Set $\underline{\mathbf{x}}_i^{(a)\prime} = (x_{i1}^{(2)}, 0.01, x_{i3}^{(1)}, x_{i4}^{(1)}, x_{i5}^{(1)}, \ldots, x_{is}^{(1)})$, find optimal α and $\underline{\mathbf{x}}_i^{(3)} = \underline{\mathbf{x}}_i^{(2)} + \alpha(\underline{\mathbf{x}}_i^{(a)} - \underline{\mathbf{x}}_i^{(2)})$.

4) Set $\underline{\mathbf{x}}_i^{(a)\prime} = (x_{i1}^{(2)}, x_{i2}^{(3)}, 0.01, x_{i4}^{(1)}, x_{i5}^{(1)}, \ldots, x_{is}^{(1)})$, find optimal α and $\underline{\mathbf{x}}_i^{(4)} = \underline{\mathbf{x}}_i^{(3)} + \alpha(\underline{\mathbf{x}}_i^{(a)} - \underline{\mathbf{x}}_i^{(3)})$.

5) Set $\underline{\mathbf{x}}_i^{(a)\prime} = (x_{i1}^{(2)}, x_{i2}^{(3)}, x_{i3}^{(4)}, 0.01, x_{i5}^{(1)}, \ldots, x_{is}^{(1)})$, find optimal α and $\underline{\mathbf{x}}_i^{(5)} = \underline{\mathbf{x}}_i^{(4)} + \alpha(\underline{\mathbf{x}}_i^{(a)} - \underline{\mathbf{x}}_i^{(4)})$.

etc.

s+1) Set $\underline{\mathbf{x}}_i^{(a)\prime} = (x_{i1}^{(2)}, x_{i2}^{(3)}, x_{i3}^{(4)}, x_{i4}^{(5)}, \ldots, x_{is-1}^{(s)}, 0.01)$, find optimal α and $\underline{\mathbf{x}}_i^{(s+1)} = \underline{\mathbf{x}}_i^{(s)} + \alpha(\underline{\mathbf{x}}_i^{(a)} - \underline{\mathbf{x}}_i^{(s)})$.

s+2) Go to (2).

Note that classification error can never increase from one step to the next. This is true because setting $\alpha = 0$ preserves the current value of classification. This process converges very quickly (usually less than 10 iterations through steps 2 to s+1 above) to a vector of coordinates which is a local maximum in terms of classification. That is, it converges to a point such that $\alpha = 0$ for all s dimensions.

In practice, the starting estimate, $\underline{\mathbf{x}}_i^{(1)}$, and the second point, $\underline{\mathbf{x}}_i^{(a)}$, could be placed anywhere within the s dimensional unit hypersphere. In practical applications the starting estimate will not be at the origin; rather, realistic starting estimates for the $\underline{\mathbf{x}}_i^{(1)}$'s will be

generated by a least squares procedure such as eigenvector/eigenvector decomposition or OLS (see Section 5 below). If the line through $\underline{\mathbf{x}}_i^{(h)}$ and $\underline{\mathbf{x}}_i^{(a)}$ is parallel to a cutting line then the corresponding difference between $w_{ij}^{(a)}$ and $w_{ij}^{(h)}$, $w_{ij}^{(a)} - w_{ij}^{(h)}$, which is used in Table 4 to find α_j, may be equal to zero. This presents no problem since if the line through $\underline{\mathbf{x}}_i^{(h)}$ and $\underline{\mathbf{x}}_i^{(a)}$ is parallel to a cutting line then the classification on that roll call is the same no matter where on the line $\underline{\mathbf{x}}_i^{(h+1)}$ is located. Consequently, the roll call is not used to locate $\underline{\mathbf{x}}_i^{(h+1)}$. In addition, if the line through $w_{ij}^{(a)}$ and $w_{ij}^{(h)}$ goes through the hypersphere so that it never intersects a cutting plane this can result in a value of α_j that produces a point that lies outside the unit hypersphere. This is easily handled by computing the upper and lower feasible limits of $\underline{\mathbf{x}}_i^{(h+1)}$ – that is, the values corresponding to the two exit points of the line from the unit hypersphere – and discarding all the corresponding L_{ij} and U_{ij} . This requires some bookkeeping but it has no effect on the search process. Finally, the search process does not have to be done by moving orthogonally (i.e., city-block-wise) through the hypersphere. However, I found it to be the most efficient way to proceed.

To guard against local maxima, I utilize multiple starting points for the $\underline{\mathbf{x}}_i^{(1)}$'s. If different solutions are found, I then search along the lines joining the unique local maxima for the best solution. After considerable experimentation, I found that 3 starting points work very well in practice. One starting point is generated from a least squares procedure and the other two are randomly generated.

Table 6 shows a Monte-Carlo study of the legislator procedure using perfect data – the true cutting planes are known -- for 100 legislators and 500 roll calls in 2 through 10 dimensions. In order to make the test of the legislator procedure reasonably stringent, I use only "unreasonable" starting points – namely, the origin and two randomly generated points are used

for the three starting points. Results for one dimension are not shown since classification will always be 100% if perfect data is used. The 100 legislators and 500 normal vectors were randomly drawn from a uniform distribution through the unit hypersphere. The cutting points along the projection line were also randomly drawn but in such a way so as to produce an average majority margin of about 62 percent (typical of congressional roll call data).[12] A maximum of 15 iterations through steps (2) - (s+1) above were allowed.

Table 6

Monte-Carlo Tests of Legislator Procedure
100 Legislators and 500 Votes

Number of Dimensions	Trials	Average Number of Errors[a]	Average Percent Correctly Classified[b]	Average Correlation True vs. Reproduced[c]
2	10	0	100.00	.982
3	10	.7	99.999	.991
4	10	.7	99.999	.989
5	10	3.2	99.999	.991
6	10	5.6	99.99	.987
7	10	9.9	99.98	.988
8	10	19.5	99.96	.985
9	10	24.8	99.95	.980
10	10	45.6	99.91	.978

[a] Average across all 10 trials.
[b] Average across all 10 trials or 500,000 total choices.
[c] The Pearson Correlation is computed between the 4950 unique legislator pair-wise distances (100x99/2). The number in the Table is the average across all 10 trials.

The legislator procedure works very well – especially at 7 dimensions and below. There is some deterioration in accuracy at 10 dimensions but it still only makes an average of 46 misclassifications out of 50,000 total choices. For 5 dimensions and below it is practically perfect. Table 6 also shows the Pearson correlation between the true configuration of legislators and the reproduced configuration. These correlations are very high. Even though the legislator procedure is non-parametric, with 500 roll call cutting planes, the unit hypersphere is chopped up into enough regions that, in effect, metric (i.e., ratio scale) information is being extracted from the roll call matrix.

5. Non-Parametric Unfolding of Binary Choice Matrices

To restate the problem, given a p by q matrix of binary choice data, **T,** the classification problem consists of finding a set of legislator points in an Euclidean space of s dimensions – the p by s matrix **X** -- and a set of cutting planes -- the q by s matrix, **N**, of normal vectors -- such that the predicted choices match the actual choices as closely as possible. To unfold the choice matrix, **T**, into the legislator coordinates and roll call cutting planes, requires a solution for two subproblems: given **X**, find the optimal **N**; and given **N**, find the optimal **X**. Sections 3 and 4 show solutions for these two subproblems. I now link them together to unfold binary choice matrices. In this section I show the algorithm and Monte-Carlo tests, and in Section 6 I show empirical examples of the algorithm.

The non-parametric unfolding algorithm consists of three phases:

1) Generate starting values for **X**, **X***, from an eigenvalue/eigenvector decomposition of the legislator by legislator agreement score matrix.
2) Given **X***, find the optimal estimate of **N**, **N***.

3) Given **N***, find the optimal **X***.

4) Go to (2).

Starting values for **X** are easily generated from an agreement score matrix. The agreement score for a pair of legislators is simply the ratio of the number of roll calls on which they voted for the same alternative (Yea and Yea, or Nay and Nay), divided by the total number of roll calls on which they both voted. This score – which ranges from 0 to 1 -- can be treated as an inverse distance – the higher the agreement score the smaller the distance between the pair of legislators. By subtracting the agreement score from 1.0 and squaring the result, the transformed agreement scores can be treated as squared distances. Double centering this matrix of squared distances – that is, from each element of the matrix subtract the row mean, subtract the column mean, and add the matrix mean – produces a cross product matrix which can be decomposed to yield an estimate of the legislator coordinates, **X*** (Young and Householder, 1938; Ross and Cliff, 1964).

Table 7 shows a Monte-Carlo study of the non-parametric unfolding algorithm using perfect data for 100 legislators and 500 roll calls in 1 through 10 dimensions. Only roll calls with margins of 97-3 to 50-50 were used because unanimous and near-unanimous roll calls trivially inflate the number of correct classifications.[13] A maximum of 25 iterations through steps (2) and (3) above were allowed. I show results for one dimension because I do not have a proof that, with perfect data, the algorithm will always converge to the true ordering of legislators and roll call midpoints. With perfect data, given the true ordering of legislators, the true midpoints are always found; and given the true ordering of midpoints, the true ordering of the legislators is always found. For the algorithm to not converge to the true joint ordering of legislators and midpoints, it must be the case that there exists a local maximum in classification.

That is, an ordering of legislators not equal to the true ordering and an ordering of midpoints not equal to the true ordering, both of which reproduce each other. Such local maxima appear to be very rare.[14]

Table 7

Monte-Carlo Tests: Non-Parametric Unfolding of Binary Choice Matrices
100 Legislators and 500 Votes (25 Iterations)

Number of Dimensions	Trials	Average Majority Margin[a]	Average Number of Errors[b]	Average Percent Correctly Classified[c]	Average Correlation True vs Reproduced[d,e]
1	10	60.3	0	100.00	1.000
2	10	61.3	27.5	99.95	.929
3	10	62.1	21.0	99.96	.961
4	10	62.4	16.5	99.97	.953
5	10	62.6	16.0	99.97	.943
6	10	63.0	11.5	99.98	.923
7	10	62.7	16.5	99.97	.913
8	10	62.7	14.5	99.97	.889
9	10	62.8	16.0	99.97	.882
10	10	62.8	13.0	99.97	.867

[a] Average margin across all 10 trials or 5000 total votes.
[b] Average across all 10 trials.
[c] Average across all 10 trials or 500,000 total choices.
[d] For s=1, the Spearman Rank Correlation is computed between the 100 true and reproduced legislator ranks. The number in the Table is the average across all 10 trials.
[e] For s > 1 the Pearson Correlation is computed between the 4950 unique legislator pair-wise distances (100x99/2). The number in the Table is the average across all10 trials.

The algorithm works well regardless of the number of dimensions. The worst result is for two dimensions where, on average, about 28 of 50,000 choices were misclassified. The accuracy of the recovery of the true configuration of legislators declines after 3 dimensions. This is to be expected given that the number of roll call cutting planes is fixed. Simply put, as the number of dimensions is increased, ceteris paribus, there is somewhat more "wiggle room" for the legislator points. However, this statistic – the correlation between the true and reproduced distances between pairs of legislators -- over-states the decline in accuracy because, if a handful of points are recovered some distance from their true location, their distances to the remaining points will deflate the correlation disproportionately. Figure 6 illustrates this point.

Figure 6A. Recovery of Legislators
True Configuration (First 35)

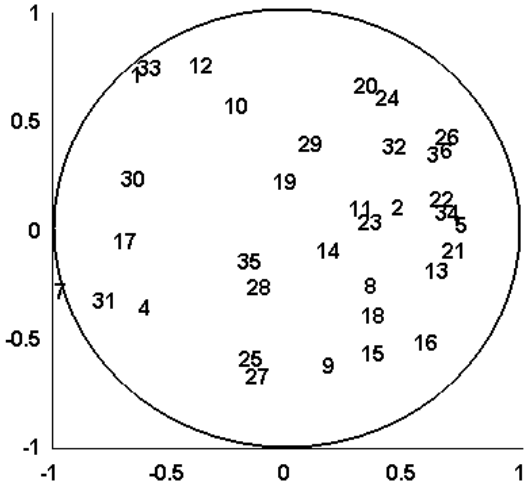
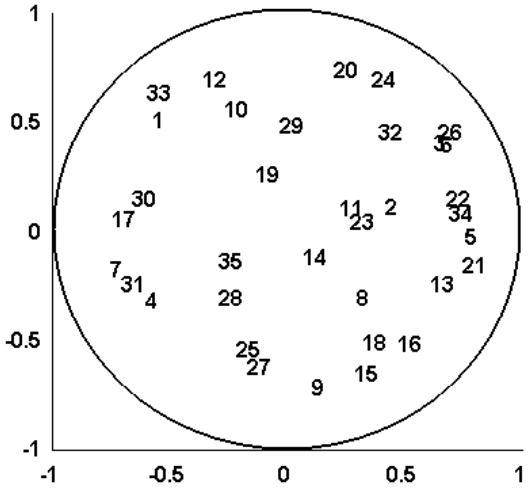


Figure 6B. Recovery of Legislators
Recovered Configuration (First 35)



The top panel of Figure 6 shows the first 35 true legislator points of one of the two-dimensional trials from Table 7 and the bottom panel shows their recovered locations.[15] (Only the first 35 are shown in order to reduce clutter.) The correlation between the 4950 unique legislator pair-wise true and reproduced distances is .928. However, the Pearson correlations between the 100 legislator positions on the corresponding first dimension is .967 and

for the second dimension the correlation is .989. The reason why the separate dimension correlations are higher than the pair-wise distance correlation can be seen in Figure 6. Note that true points 7, 1, and 33 are near the rim of the circle but are recovered somewhat to the interior. However, the relative placement of all the points in the recovered configuration is very close to the true. This distortion of the points near the rim has a greater impact on the distance correlation than it does on the dimension correlations.

Given the history of other multidimensional scaling techniques, most empirical applications of the non-parametric unfolding technique I show here will be to data matrices with missing entries and the estimated configurations will be in three or fewer dimensions.[16] Missing data presents no problem for the algorithm. In the cutting plane procedure it simply means that the total number of legislators may vary from vote to vote. In the legislator procedure it simply means that the number of cutting lines may vary from legislator to legislator. Handling missing data requires a little bookkeeping but it has no effect on the algorithm.

Table 8 shows a set of experiments with binary choice data with and without error at various levels of missing data. Configurations of 100 legislators and 500 roll calls in 2 and 3 dimensions were randomly generated in the same fashion as those used in the Monte-Carlo experiments shown in Table 7. Error was introduced into the choices by making them probabilistic where the probability of making a correct choice increases with distance from the cutting plane.[17] An error level of about 22 percent was chosen because that is the approximate level of error in U.S. congressional roll call data. Entries were randomly removed and the remaining entries were then analyzed by the algorithm in one through five dimensions.[18] The upper part of Table 8 shows two-dimensional experiments at four different levels of missing data with and without error, and the lower part shows three-dimensional experiments. Each

randomly produced matrix was analyzed at each level of missing data so that the same 10 matrices for two or three dimensions (with varying levels of missing entries) are being averaged in each row of the upper or lower parts of the Table.

Table 8

Monte-Carlo Tests: Non-Parametric Unfolding of Binary Choice Matrices With Missing Data (Each Entry Average of 10 Trials, 15 Iterations Per Trial)

2 Dimensions, 100 Legislators, 500 Votes

Percent Missing	Average Percent Error	Average Majority Margin	Percent Correct 1 Dim.	Percent Correct 2 Dim.	Percent Correct 3 Dim.	Percent Correct 4 Dim	Percent Correct 5 Dim.	R All	R 1st	R 2nd
0	0	61.7	90.1	99.9	100.0	100.0	100.0	.942	.982	.982
20	0	62.2	90.3	99.9	100.0	100.0	100.0	.943	.985	.982
50	0	62.6	90.1	99.9	100.0	100.0	100.0	.932	.982	.978
70	0	64.1	90.5	99.8	99.9	100.0	100.0	.903	.974	.966
0	21.1	61.1	77.0	82.3	83.0	83.6	84.2	.940	.983	.982
20	21.1	61.5	77.5	83.0	83.6	84.3	85.0	.936	.981	.980
50	21.1	62.3	78.4	84.2	85.3	86.2	86.9	.922	.976	.973
70	21.1	63.6	79.6	86.1	87.6	88.9	90.1	.888	.963	.961

3 Dimensions, 100 Legislators, 500 Votes

Percent Missing	Average Percent Error	Average Majority Margin	Percent Correct 1 Dim.	Percent Correct 2 Dim.	Percent Correct 3 Dim.	Percent Correct 4 Dim	Percent Correct 5 Dim.	R All	R 1st	R 2nd	R 3rd
0	0	62.2	84.1	91.5	99.9	100.0	100.0	.958	.992	.992	.990
20	0	62.5	84.3	91.5	99.9	100.0	100.0	.937	.990	.991	.988
50	0	63.0	84.6	91.8	99.8	99.9	100.0	.874	.985	.985	.980
70	0	64.5	84.6	92.6	99.7	99.8	99.9	.786	.975	.974	.965
0	22.8	61.3	73.8	77.7	81.2	81.9	82.6	.921	.982	.980	.979
20	22.8	61.7	74.1	78.3	81.9	82.6	83.3	.915	.981	.977	.974
50	22.8	62.4	74.9	79.9	83.4	84.3	85.3	.872	.969	.966	.952
70	22.8	63.4	76.4	82.1	85.6	87.1	88.5	.784	.947	.938	.908

The accuracy of the recovery of the legislator configuration is quite good and only begins to fall off at 70 percent missing entries. Consistent with the discussion of Figure 6, the overall Pearson correlation between the 4,950 pairwise true and reproduced distances deteriorates more rapidly than the correlations between the true and reproduced legislator locations on each dimension.

Figure 7 shows the average percent correct classifications from Table 8. The elbows are quite clear at the true dimensionality. With perfect data the procedure unambiguously finds the true dimensionality. With error, there is a tendency for the correct classification to increase with the percentage of missing data. This makes sense because, with respect to locating a legislator point, with more missing data there are fewer roll call cutting planes and hence the legislator position is not as constrained as it is with complete data. This increase in "wiggle room" will increase the correct classification and decrease the correlations of the true and reproduced legislator configurations. In any event, the results shown in Table 8 and Figure 7 suggest that the algorithm will perform well with real world data at realistic levels of missing entries.

Figure 7A. Two Dimensions
Perfect Data With Missing Entries

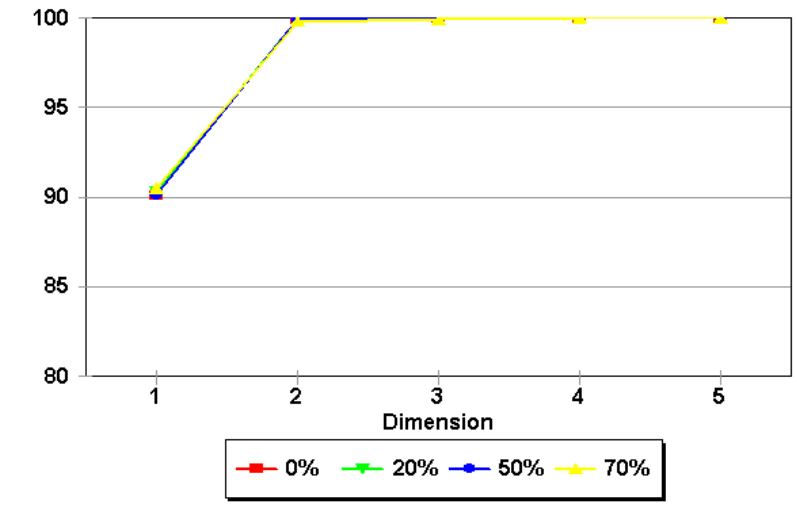


Figure 7B. Two Dimensions
Error With Missing Entries

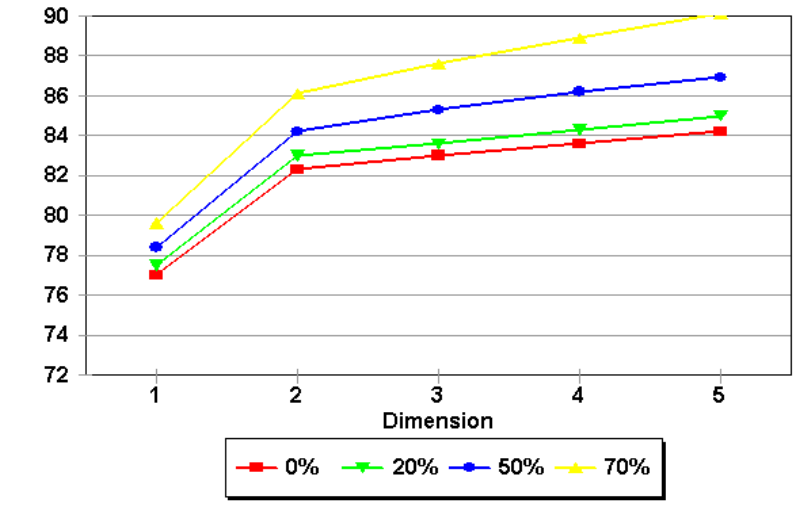


Figure 7C. Three Dimensions
Perfect Data With Missing Entries

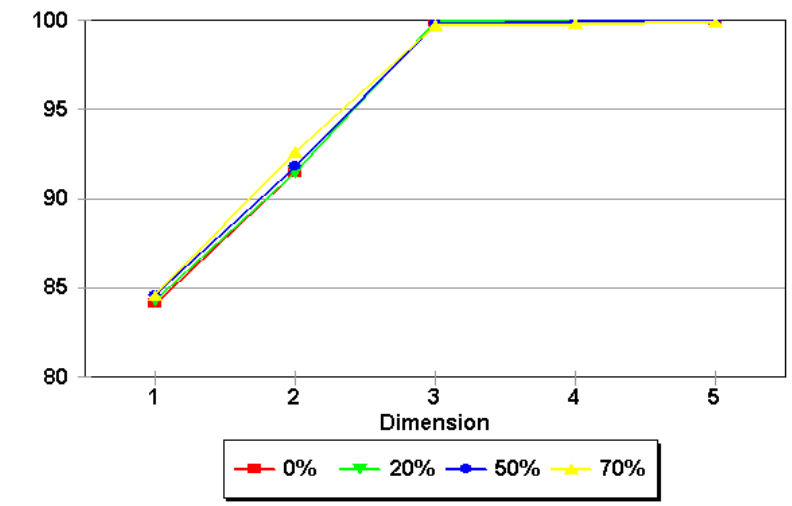
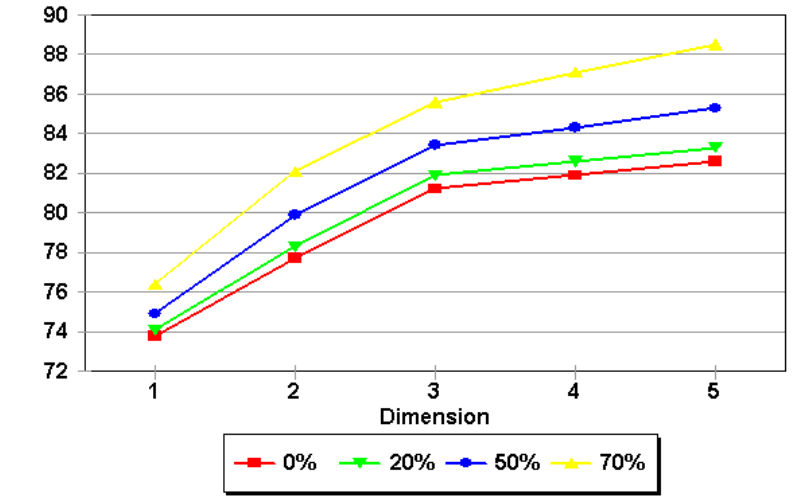


Figure 7D. Three Dimensions
Error With Missing Entries



6. Empirical Examples of Non-Parametric Unfolding of Binary Choice Matrices

In this section I show two empirical examples of the non-parametric unfolding procedure. The first is to U.S. Senate roll call data and the second is to feeling thermometer ratings of political figures gathered from respondents in the NES 1968 presidential election study. Both

sets of data have been extensively analyzed by numerous researchers utilizing a variety of methodologies.

Feeling thermometer data is technically not binary choice – however, it can be interpreted as rank order data and that can be converted to binary choice. A feeling thermometer measures how warm or cold a person feels towards the stimulus and the measure ranges from 0 – very cold and unfavorable opinion – to 100 – very warm and favorable opinion with 50 being a neutral point. In 1968 respondents were asked to give feeling thermometer ratings to 12 political figures: George Wallace, Hubert Humphrey, Richard Nixon, Eugene McCarthy, Ronald Reagan, Nelson Rockefeller, Lyndon Johnson, George Romney, Robert Kennedy, Edmund Muskie, Spiro Agnew, and Curtis LeMay.[19]

Suppose a respondent gave ratings of 30, 80, and 55 to Wallace (W), Humphrey (H), and Nixon (N) respectively. With respect to these three candidates, the rank order is H > N > W. Now suppose a second respondent gave ratings of 45, 65, and 95, respectively, for a rank order of N > H > W. These rank orders can be converted to binary choice data by treating each pair of candidates as a roll call vote. For example, consider the pair of Wallace and Humphrey. If a respondent rates Wallace higher than Humphrey make that Yea, and if Humphrey is rated higher than Wallace, make that Nay. Doing this consistently across respondents creates a roll call vote where the outcomes are Wallace and Humphrey, respectively.

With the actual 1968 data, I used the order of the 12 political figures listed above (which is their actual order in the NES data set) to create the roll calls. That is, given a pair of politicians, the one earlier in the NES ordering was treated as a Yea and the later one a Nay. So if the pair was Ronald Reagan and Curtis LeMay, then if a respondent rated Reagan higher than LeMay that is a Yea vote. If a respondent gave a pair of politicians the same rating, for

example, 55 and 55, then I treated it as missing data (that is, as if the respondent abstained on the roll call).

6.a. U.S. Senate Roll Call Matrices

My first application is to Senate voting after World War II. I focus on this period because it has been extensively analyzed by Poole and Rosenthal (1997) which will facilitate the interpretation of the results. I compare the two-dimensional senator coordinates from the non-parametric unfoldings with those produced by KYST, a non-metric multidimensional scaling procedure developed by Kruskal, Young, and Seery (1973), and NOMINATE, a maximum likelihood procedure developed by Poole and Rosenthal (1991, 1997).

Table 9 reports the classification results for Senates 80 to 104 (first session) in one and two dimensions for the non-parametric procedure. These percentages are about 3 to 5 percentage points better than NOMINATE in both one and two dimensions (Poole and Rosenthal, 1997, chapter 3). This is not surprising given that the NOMINATE procedure maximizes a likelihood function – that is, it estimates legislator and roll call outcome coordinates which maximize the probabilities of the observed choices. It does not attempt to maximize correct classifications.

Table 9

U.S. Senate: 1947 - 1995
Non-Parametric Unfolding of Roll Call Data

Senate	Years	Senators	Roll Calls	Total Choices	Average Margin	Non-P 1st	Non-P 2nd	kyst R 1st	kyst R 2nd	nom R 1st	Nom R 2nd
104	1995	101[a]	541[b]	52,966[c]	.638	90.1[d]	91.4	.977[e]	.605	.976[f]	.852
103	1993-94	101	647	63,023	.672	89.2	90.4	.988	.884	.994	.943
102	1991-92	102	481	46,208	.685	86.9	88.5	.993	.899	.981	.926
101	1989-90	101	499	48,649	.680	85.4	87.1	.992	.919	.987	.949
100	1987-88	101	635	59,631	.709	87.7	89.5	.991	.854	.991	.971
99	1985-86	101	661	63,104	.688	84.7	86.8	.996	.917	.992	.978
98	1983-84	101	578	53,330	.698	84.8	87.3	.993	.940	.989	.975
97	1981-82	101	818	77,672	.682	85.5	88.1	.997	.948	.996	.987
96	1979-80	101	928	82,937	.683	83.5	85.8	.994	.895	.995	.984
95	1977-78	104	1037	92,868	.691	84.5	86.4	.996	.844	.993	.737
94	1975-76	100	1144	100,328	.691	86.3	88.6	.995	.982	.993	.981
93	1973-74	101	983	87,699	.695	85.1	87.5	.997	.953	.997	.977
92	1971-72	102	783	68,588	.676	85.0	88.6	.995	.963	.993	.983
91	1969-70	102	557	49,219	.681	84.5	88.1	.995	.951	.987	.968
90	1967-68	101	518	46,081	.699	83.6	87.2	.992	.949	.993	.968
89	1965-66	102	441	40.618	.681	85.4	88.4	.991	.949	.983	.974
88	1963-64	102	505	47,797	.686	85.0	90.1	.976	.978	.936	.965
87	1961-62	105	400	38,189	.675	87.3	90.6	.975	.971	.953	.923
86	1959-60	103	360	33,855	.686	84.9	89.6	.981	.971	.971	.954
85	1957-58	98	255	23,097	.669	84.7	89.4	.979	.963	.961	.959
84	1955-56	99	184	16,798	.659	85.5	90.4	.982	.955	.955	.960
83	1953-54	103	242	20,991	.672	86.9	90.3	.974	.859	.956	.938
82	1951-52	96	208	17,368	.659	86.0	89.4	.971	.848	.975	.938
81	1949-50	102	447	38,074	.667	85.0	88.6	.985	.925	.979	.963
80	1947-48	97	237	20,321	.665	88.0	90.8	.967	.932	.969	.903

[a] Number of Senators may exceed two times the number of States because of within Congress replacements.
[b] Number of roll calls with at least 2.5% voting, paired, or announced, on losing side.
[c] Total choices may not equal number of Senators times number of roll calls because of non-voting due to absences, etc..
[d] Classifications from non-parametric unfolding algorithm.
[e] Pearson correlation between Senator coordinates from KYST and Senator coordinates from non-parametric unfolding. Non-parametric unfolding configuration rotated to best match KYST configuration.
[f] Pearson correlation between Senator coordinates from W-NOMINATE and Senator coordinates from non-parametric unfolding. Non-parametric unfolding configuration rotated to best match W-NOMINATE configuration.

Table 9 also shows the Pearson correlations between the estimated dimensions of the non-parametric procedure and those produced by KYST and NOMINATE, respectively, in two dimensions. [20] These correlations are, for the most part, very high – most of the first dimension correlations are above .95 and the second dimension correlations are mostly above .9. Because the non-parametric unfolding produces configurations very similar to those of both KYST and NOMINATE, these results strongly support the substantive interpretations of the legislator configurations discussed in Poole and Rosenthal (1997).

Table 10 shows the rank order from most liberal to most conservative for the 104th Senate using roll call data through December 1995. Campbell of Colorado switched from Democrat to Republican in April of 1995 so he appears twice (ranks 47 and 53). If two or more senators tied in the ranking, the average of the associated ranks was used. For example, 72 senators were more liberal and 26 more conservative than the threesome Shelby (R-AL), Abraham (R-MI), and Frist (R-TN), who were tied. Consequently they all were assigned the average rank of 74.

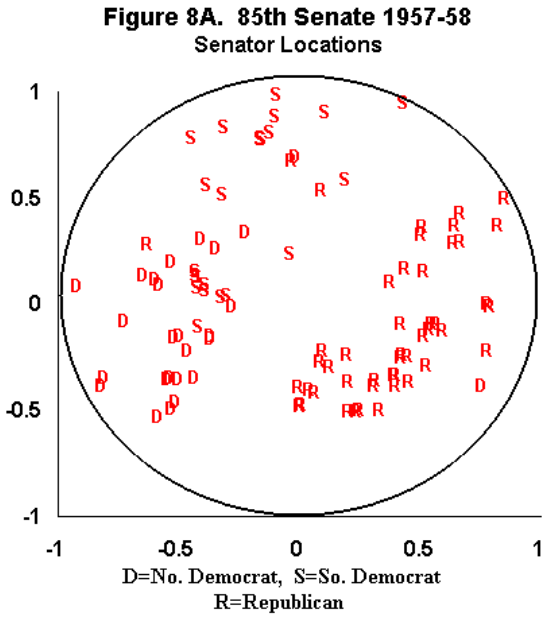
Table 10

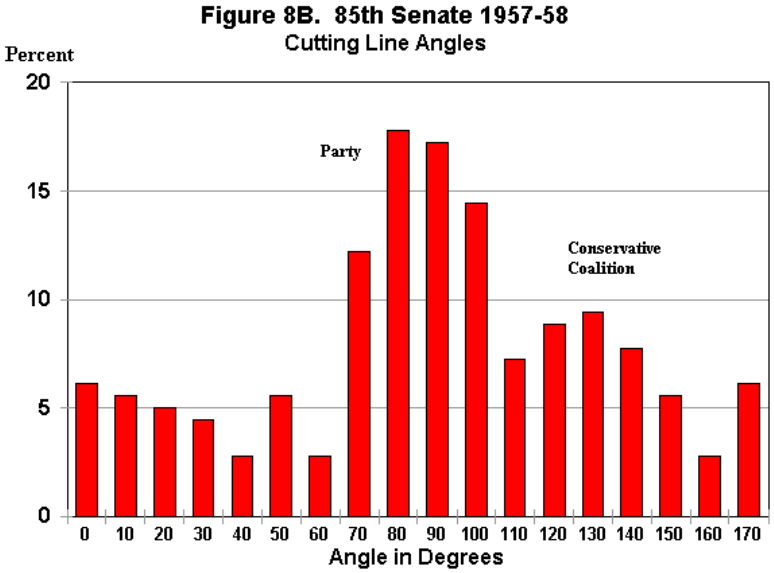
104th (1995) U.S. Senate

Name	Rank	Name	Rank	Name	Rank
Simon (D-IL)	1	Nunn (D-GA)	45	Lott (R-MS)	89
Wellstone (D-MN)	2	Heflin (D-AL)	46	Gramm (R-TX)	90
Feingold (D-WI)	3	Campbell (D-CO)	47	Helms (R-NC)	91
Levin (D-MI)	4	Jeffords (R-VT)	48	Craig (R-ID)	92
Kennedy (D-MA)	5	Cohen (R-ME)	49	Kempthorne (R-ID)	93
Boxer (D-CA)	6	Specter (R-PA)	50	Nickles (R-OK)	94
Leahy (D-VT)	7	Snowe (R-ME)	51	Smith (R-NH)	95.5
Bumpers (D-AR)	8	Chafee (R-RI)	52	Inhofe (R-OK)	95.5
Bradley (D-NJ)	9.5	Campbell (R-CO)	53	Faircloth (R-NC)	97
Lautenberg (D-NJ)	9.5	Kassebaum (R-KS)	54	Grams (R-MN)	98
Murray (D-WA)	11	Packwood (R-OR)	55	Brown (R-CO)	99
Harkin (D-IA)	12	Simpson (R-WY)	56	Kyl (R-AZ)	100.5
Moseley-Braun (D-IL)	13	Roth (R-DE)	57	McCain (R-AZ)	100.5
Pell (D-RI)	14	Hatfield (R-OR)	58		
Moynihan (D-NY)	15	Dewine (R-OH)	59		
Dorgan (D-ND)	16	Stevens (R-AK)	60		
Conrad (D-ND)	17	Gorton (R-WA)	61		
Pryor (D-AR)	18	D'Amato (R-NY)	62		
Kerry (D-MA)	19.5	Domenici (R-NM)	63		
Kohl (D-WI)	19.5	Lugar (R-UT)	64		
Sarbanes (D-MD)	21	Bond (R-MO)	65		
Akaka (D-HI)	22	Pressler (R-SD)	66		
Daschle (D-SD)	23	Murkowski (R-AK)	68		
Rockefeller (D-WV)	24	Cochran (R-MS)	68		
Biden (D-DE)	25	Burns (R-MT)	68		
Mikulski (D-MD)	26	Warner (R-VA)	70		
Dodd (D-CT)	27	Grassley (R-IA)	71		
Glenn (D-OH)	28	Thomas (R-WY)	72		
Inouye (D-HI)	29	Shelby (R-AL)	74		
Bingaman (D-NM)	30	Abraham (R-MI)	74		
Byrd (D-WV)	31	Frist (R-TN)	74		
Bryan (D-NV)	32	Hatch (R-UT)	76		
Graham (D-FL)	33	Bennett (R-UT)	77		
Feinstein (D-CA)	34	Santorum (R-PA)	78		
Kerrey (D-NE)	35	Hutchison (R-TX)	79		
Ford (D-KY)	36	Gregg (R-NH)	80		
Hollings (D-SC)	37	Mack (R-FL)	81.5		
Reid (D-NV)	38	Dole (R-KS)	81.5		
Breaux (D-LA)	39	Coverdell (R-GA)	83.5		
Johnston (D-LA)	40	Coats (R-IN)	83.5		
Robb (D-VA)	41	McConnell (R-KY)	85.5		
Lieberman (D-CT)	42	Thurmond (R-SC)	85.5		
Exon (D-NE)	43	Thompson (R-TN)	87		
Baucus (D-MT)	44	Ashcroft (R-MO)	88		

The polarization of American politics is evident from an inspection of the table. There is no overlap of the two parties.[21] Campbell's voting record as a Democrat made him the most conservative Democrat in the Senate. His conversion only moved him from 47th to 53rd rank – from the right edge of the Democratic party to the midst of the moderates of the Republican party.

Figure 8 shows the two dimensional configuration of senators for the 85th Senate along with a histogram of the roll call cutting line angles. The two major parties are clearly separated with the Democratic party being split into its Northern and Southern wings. The 85th Senate occurred during the height of the three-party system which lasted from the late 1930s to the late 1970s (Poole and Rosenthal, 1991, 1997; McCarty, Poole, and Rosenthal, 1996). The approximate angle of a party-line vote and the approximate angle of a conservative coalition vote (Northern Democrats versus a coalition of Southern Democrats and Republicans) are indicated in the histogram of the cutting line angles. The second dimension picked up the split in the Democratic Party over race-related issues.





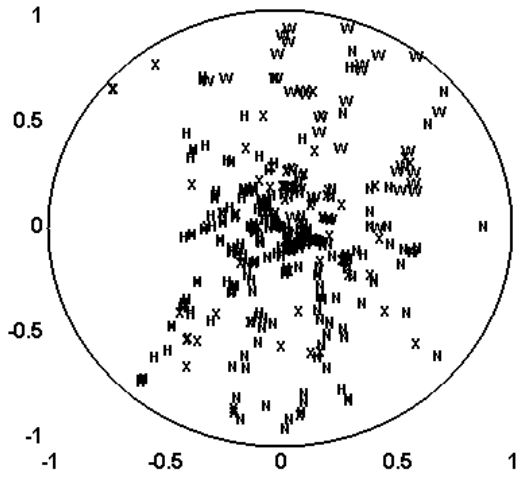
6.b. NES 1968 Feeling Thermometer Ratings of Political Figures

The thermometer ratings were converted to binary choices as explained earlier. If a respondent gave a thermometer rating to all 12 politicians, then the respondent was treated as casting 66 (12x11/2) roll call votes. In order to be included in the analysis, the respondent had

to vote on at least 25 out of the 66 total possible pairings. The first 450 respondents in the survey were analyzed of which 418 rated enough politicians to be included in the analysis.[22]

The results were quite good. The correct classifications were 84.6 percent in one dimension and 88.9 percent in two dimensions (average margin, 0.699; total choices, 21,839). Figure 9 shows several plots of the 418 respondents coded as to how they reported they had voted (or not voted) in the 1968 presidential election.

Figure 9A. 1968 Voters and Non-Voters



X=Non-Voter, H=Humphrey
N=Nixon, W=Wallace

Figure 9B. 1968 Non-Voters

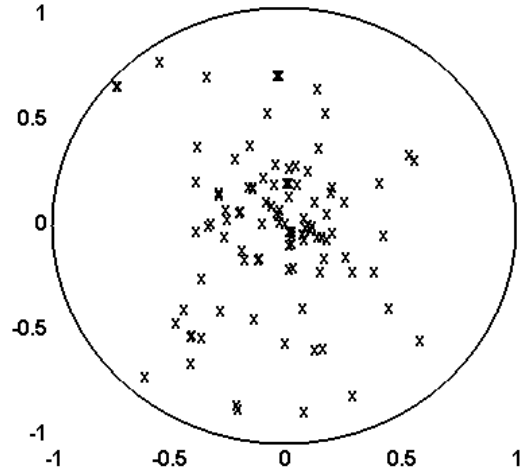


Figure 9C. 1968 Humphrey Voters

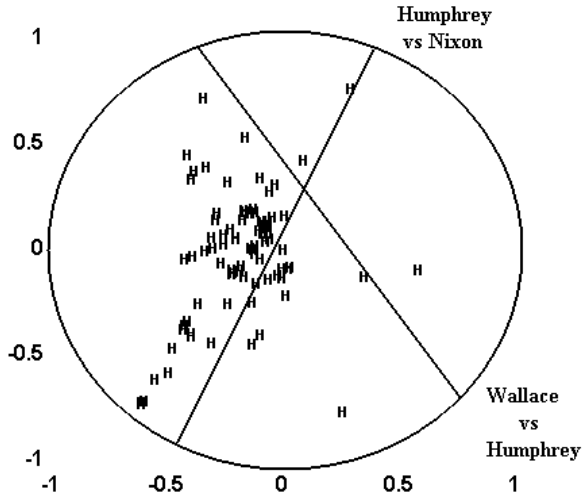


Figure 9D. 1968 Nixon Voters

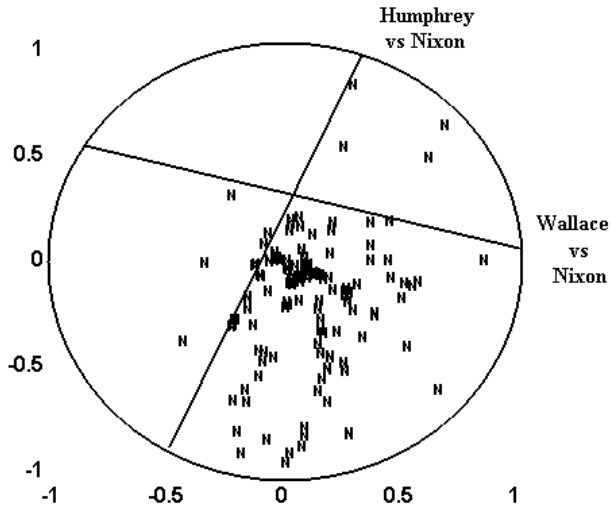
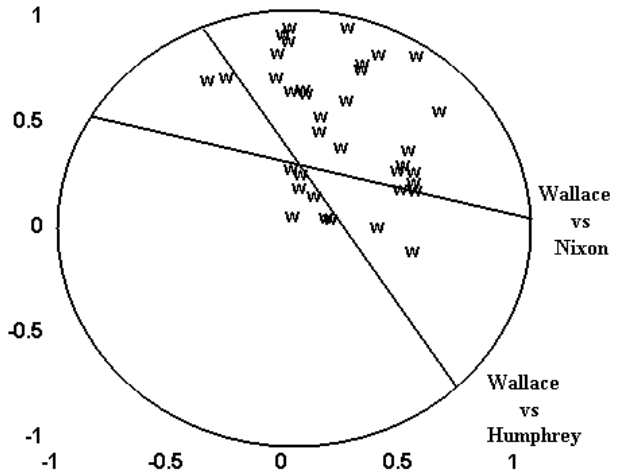


Figure 9E. 1968 Wallace Voters



Panel A of Figure 9 shows all 418 respondents, panel B just the non-voters, and panels C, D, and E show the Humphrey, Nixon, and Wallace voters, respectively, along with the cutting lines between the relevant presidential candidate pairs. The distribution of the voters and non-voters is very similar to that found by other scaling techniques (Wang et al. [1975]; Rabinowitz [1976]; Cahoon et al. [1978]; Poole and Rosenthal [1984]).[23] Specific locations for the political figures cannot be estimated with this technique. However, ceteris paribus, the cutting lines do create regions where the politicians must lie. For example, there are no cutting lines for Wallace that are above the two shown in panel E. The cutting lines for Wallace versus Johnson and Wallace versus Kennedy are almost exactly where the Wallace versus Humphrey line is in the Figure. Hence, Wallace must be located in the pie slice defined by the Wallace versus Nixon and Wallace versus Humphrey cutting lines.

Wallace was such a divisive figure in 1968 that the cutting lines for Wallace versus the other 11 political figures all classify at 95 percent or higher. Consequently, the noise level is so low that the region of the space that Wallace must lie in can be inferred with some confidence.

This is not as true for either Humphrey or Nixon. For example, when paired against Robert Kennedy, 78 respondents voted for Humphrey and 223 voted for Kennedy. The estimated cutting line produced 66 classification errors – little better than the marginals of the roll call. Consequently, the cutting line for Humphrey versus Kennedy yields little useful information. However, this is an exception, not the rule. The overall correct classification in two dimensions was 88.9 percent and most of the cutting lines for Nixon and Humphrey are well above that figure. Although I will not pursue the topic here, it should be possible to infer what regions the outcomes are in when faced with noisy cutting lines.

7. Conclusion

In this paper I have shown a general non-parametric technique for maximizing the correct classification of binary choice or two-category data. I estimate cutting planes or cutting planes and chooser points in an Euclidean space such that the correct classification of the observed two-category data is maximized. I make only two assumptions: 1) the choice space is Euclidean; and 2) the individuals making choices behave as if they utilize symmetric, single-peaked preferences. I strongly suspect that the first assumption can be relaxed to a general Minkowski metric. That is a topic for future research. However, the assumption of symmetric preferences cannot be relaxed.

In order to perform a non-parametric unfolding of a binary choice matrix, two subproblems must be solved. First, given the chooser coordinates, for each binary choice find the cutting plane that maximizes correct classification; and second, given the cutting planes, for each chooser find the point that maximizes correct classification. Solutions for these two subproblems were shown in sections 3 and 4.

Although I do not have formal proofs that either technique converges to the classification maximum, Monte-Carlo tests show that both in fact work very well in practice. In the presence of error, the cutting plane procedure does not necessarily converge to a classification maximum. However, because of the way that the cutting plane procedure is operationalized, it almost certainly passes through, or very near to, the classification maximum and the maximum can be recovered from the iteration record. The legislator/chooser procedure is guaranteed to converge to a very strong local maximum. That is, a local maximum for which the point cannot be moved in any orthogonal direction and have the correct classifications increase. When the two procedures are used together in an alternating framework to analyze binary choice matrices, their performance is very good. The Monte-Carlo tests in section 5 and the empirical applications in section 6 are testimony to this fact.

For data sets which consist of a two-category dependent variable and a set of independent variables, the cutting plane procedure shown in section 3 recovers essentially the same coefficients as a probit analysis of the same data when the underlying error distribution is symmetric. The bootstrapped standard errors for the cutting plane coefficients almost always produce the same pattern of significance as that for the probit coefficients.

The cutting plane procedure can be easily generalized to multi-category ordered probit using a single normal vector. In the multiple choice context where there is no natural ordering of the choices so that there are multiple normal vectors, it should be possible to modify the cutting plane procedure to handle curved surfaces. This is a topic for future research. If successful, it would permit the non-parametric estimation of multiple choice and conditional choice models.

References

Cahoon, Lawrence S., Melvin J. Hinich, and Peter C. Ordeshook. 1978. "A Statistical Multidimensional Scaling Method Based on the Spatial Theory of Voting." In *Graphical Representation of Multivariate Data*, edited by P. C. Wang. New York: Academic Press.

Eckart, Carl and Gale Young. 1936. "The Approximation of One Matrix by Another of Lower Rank." *Psychometrika*, 1:211-218.

Green, Paul E., Frank J. Carmone Jr., and Scott M. Smith. 1989. *Multidimensional Scaling: Concepts and Applications*. Boston: Allyn and Bacon.

Johnson, Richard M. 1963. "On a Theorem Stated by Eckart and Young." *Psychometrika*, 28:259-263.

Keller, Joseph B. 1962. "Factorization of Matrices by Least-Squares." *Biometrika*, 49:239-242.

Kruskal, Joseph B. 1964a. "Multidimensional Scaling by Optimizing a Goodness of Fit to a Nonmetric Hypothesis." *Psychometrika*, 29:1-27.

Kruskal, Joseph B. 1964b. "Nonmetric Multidimensional Scaling: A Numerical Method." *Psychometrika*, 29:115-129.

Kruskal, Joseph B. and Myron Wish. 1978. *Multidimensional Scaling*. Beverly Hills, Ca: Sage Publications.

Kruskal, Joseph B., Forrest W. Young, and J. B. Seery. 1973. "How to Use KYST: A Very Flexible Program to Do Multidimensional Scaling and Unfolding." *Multidimensional Scaling Program Package of Bell Laboratories.* Bell Laboratories, Murray Hill, N.J.

McCarty, Nolan M., Keith T. Poole, and Howard Rosenthal. 1996. "The Realignment of American Politics: From Goldwater to Gingrich." Manuscript, GSIA, Carnegie-Mellon University, Pittsburgh, PA 15213.

Palfrey, Thomas R. and Keith T. Poole. 1987. "The Relationship Between Information, Ideology, and Voting Behavior." *American Journal of Political Science*, 31:511-530.

Poole, Keith T. 1990. "Least Squares Metric, Unidimensional Scaling of Multivariate Linear Models." *Psychometrika*, 55:123-149.

Poole, Keith T. and Howard Rosenthal. 1984. "U.S. Presidential Elections 1968-1980: A Spatial Analysis." *American Journal of Political Science*, 28:282-312.

Poole, Keith T. and Howard Rosenthal. 1991. "Patterns of Congressional Voting." *American Journal of Political Science,* 35:228-278.

Poole, Keith T. and Howard Rosenthal. 1997. *Congress: A Political-Economic History of Roll Call Voting*. New York: Oxford University Press.

Rabinowitz, George. 1976. "A Procedure for Ordering Object Pairs Consistent With the Multidimensional Unfolding Model." *Psychometrika,* 45:349-373.

Ross, John and Norman Cliff. 1964. "A Generalization of the Interpoint Distance Model." *Psychometrika,* 29:167-176.

Schonemann, Peter H. 1966. "A Generalized Solution of the Orthogonal Procrustes Problem." *Psychometrika,* 31:1-10.

Shepard, Roger N. 1962a. "The Analysis of Proximities: Multidimensional Scaling With an Unknown Distance Function: I." *Psychometrika*, 27: 125-140.

Shepard, Roger N. 1962b. "The Analysis of Proximities: Multidimensional Scaling With an Unknown Distance Function: II." *Psychometrika*, 27: 219-246.

Weisberg, Herbert F. and Jerrold G. Rusk. 1970. "Dimensions of Candidate Evaluation." *American Political Science Review*, 64:1167-1185.

Wang, Ming-Mei, Peter H. Schonemann, and Jerrold G. Rusk. 1975. "A Conjugate Gradient Algorithm for the Multidimensional Analysis of Preference Data." *Multivariate Behavioral Research,* 10:45-80.

Young, Gale and A. S. Householder. 1938. "Discussion of a Set of Points in Terms of Their Mutual Distances." *Psychometrika*, 3:19-22.

Notes

[1] The configuration shown in Figures 3 and 4 is of the 80th House of Representatives. See Poole and Rosenthal (1991, 1997) for an extended discussion of the analysis of roll call voting in the U.S. Congress.

[2] The famous Eckart-Young result was never stated as an explicit theorem in their paper. Rather they use two theorems from linear algebra and a very clever argument to show the truth of their result. Later, Keller (1962) independently rediscovered the Eckart-Young result.

[3] The first proof that every rectangular matrix of real elements can be decomposed as shown in equation (5), was given by Johnson (1963).

[4] In particular, each legislator coordinate was drawn from a uniform [-1,+1] distribution. If the sum of the squared coordinates exceeded 1, the coordinates were discarded and a new draw performed. In effect the draw was over the unit hypercube and those coordinates outside the internal unit hypersphere were discarded. The coordinates for the normal vectors were also drawn from a [-1,+1] distribution and then normalized so that their sum of squares equaled 1. The cutting points were drawn from a uniform [0,1] distribution and then taken to the 4th power. This had the desired effect of clustering the cutting planes such that the average margin was about 62-38 (or 38-62 since the yes/no outcomes were also randomly assigned).

[5] X was generated as explained in note 4 above.

[6] The s elements of β were randomly drawn from the uniform [-1,+1] distribution and then scaled so that their sum of squares equaled 1; that is, $\beta'\beta = 1$.

[7] The Probit classifications were based upon the estimated probabilities from the Probit analysis.

[8] As of April, 1996, there were 236 Republicans in the House. I excluded the 5 party switchers – Laughlin (TX), Parker (MS), Hayes (LA), Deal (GA), Tauzin (LA) – from the analysis. Campbell (R-CA), who won a special election to replace Mineta (D-CA), is included.

[9] The source for the co-sponsors is "Who You Calling 'Moderate?'", by Bob Balkin, PoliticsUSA, at www.politicsusa.com, Wednesday, April 24, 1996.

[10] I sampled by observation with replacement (that is, I sampled the rows of the original data matrix with replacement) to form 100 matrices and ran the cutting plane procedure on each matrix. The standard errors were obtained by computing the sum of squared differences between the actual normal vector from the original data and the 100 normal vectors from the bootstrap trials. I divided this sum of squares for each coefficient by 100 and took the square root as the estimate of the standard error.

[11] See Palfrey and Poole (1987) for a conditional logit model of choice for the 1980 presidential election.

[12] See note 4 above.

[13] The data were generated as described in note 4 above. Roll calls with less than 2.5% (98-2, 99-1, 100-0) in the minority were discarded so that the results reported here can be compared to those from the NOMINATE procedure developed by Poole and Rosenthal (1991, 1997). See Section 6 below.

[14] In parametric problems which use a squared error loss function such as one dimensional metric similarities/unfolding analysis, this sort of local minimum is extremely rare. See Poole (1990).

[15] The recovered locations of the 100 artificial legislators were rotated so as to best match the true for purposes of presentation. The rotation does not change the interpoint distances between the legislators. In psychometrics, this is known as an "orthogonal procrustes" problem. I used the technique developed by Schonemann (1966) to solve for the rotation matrix.

[16] See Kruskal and Wish (1978) for a general discussion and examples; and Green, Carmone, and Smith (1989) for numerous examples in the marketing field.

[17] To generate the probabilities, I utilized a simple logit model framework. Yea and Nay outcome coordinates were created by placing them on the normal vector .5 units on either side of the true cutpoint. A legislator's utility for each outcome was assumed to be a bell-shaped function in the squared distance (d^2) of the legislator to an outcome plus random error; that is, the utility function is

$$U(\text{Yea}) = u(\text{Yea}) + \varepsilon = \alpha\exp[\lambda d^2] + \varepsilon$$

where u(Yea) is the deterministic portion of the utility function, and ε, the stochastic portion, is distributed as the log of the inverse exponential (the logit distribution). Because the logit distribution does not have a variance parameter, α is a scaling constant that controls the overall noise level – if α is zero, the utility will be white noise, if α is large, the utility will be due only to the distance. λ is a scaling constant that controls the shape of the utility function. The actual values used were α=15.0 and λ=.125. The probability of voting Yea is: $P(\text{Yea}) = \exp[u(\text{Yea})]/\{\exp[u(\text{Yea})] + \exp[u(\text{Nay})]\}$. The choice with the higher probability was entered in the matrix.

[18] For every entry in the 100 by 500 matrix a "weighed coin" was "flipped" to determine if it was to be removed. This was accomplished by drawing 50,000 numbers from the uniform [0, 1] distribution – one for each entry in the roll call matrix. To generate 20 percent missing data, for example, all entries corresponding to a uniform random number greater than .8 are removed.

[19] The NES survey was conducted after Robert Kennedy's assassination in June, 1968. This obviously affects the ratings Kennedy received.

[20] I limit my analysis to the two dimensional case because Poole and Rosenthal (1997) show that, for most of American history, at most two dimensions are required to account for the substance of roll call voting decisions. The non-parametric configuration was rotated to best match the NOMINATE configuration using Schonemann's (1966) technique. See note 15 above.

[21] For a general analysis of the polarization of American politics, see McCarty, Poole, and Rosenthal (1996). Using NOMINATE, they show almost no overlap in the House of Representatives.

[22] There were a total of 1399 respondents. I limited the number of respondents to 450 due to computer memory considerations.

[23] Weisberg and Rusk (1970) use the non-metric multidimensional scaling procedure developed by Kruskal (1964a,b) to recover a candidate configuration from the candidate by candidate Pearson correlation matrix computed across the respondents. They do not estimate the respondent locations. The candidate configuration estimated by Weisberg and Rusk is essentially the same as that estimated by the other cited researchers.