ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/227390319

Spatial Models of Parliamentary Voting

Book · April 2005

DOI: 10.1017/CBO9780511614644 · Source: RePEc

CITATIONS
491

READS
2,513

1 author:

Keith T. Poole
University of Georgia
124 PUBLICATIONS **12,744** CITATIONS

SEE PROFILE

All content following this page was uploaded by Keith T. Poole on 21 May 2014.

The user has requested enhancement of the downloaded file.

SPATIAL MODELS OF PARLIAMENTARY VOTING

By

Keith T. Poole

Chapter 4: Probabilistic Spatial Models of Parliamentary Voting

Overview

Several probabilistic models of legislative voting have been developed during the past 30 years. In my discussion of the Optimal Classification (OC) method in Chapter 3, I assumed only that legislators used a symmetric single-peaked utility function. I did not make any assumptions about the functional form of the error distribution or the functional form of the deterministic utility function. OC is simply a non-parametric scaling method that maximizes correct classifications. The downside to OC is that it is difficult to do any meaningful *statistical inference* beyond straightforward Monte Carlo testing (Poole, 1997; 2000a; b). Determining the statistical properties of OC is an unsolved problem.

Below I make explicit assumptions about the functional form of both the deterministic utility function and the error distribution. These assumptions have consequences – there is no free lunch with probabilistic voting models. Estimation of these models is complicated, and the statistical issues are especially thorny and not fully resolved. Despite these drawbacks, the good news is that the major probabilistic methods produce essentially the same spatial maps for large legislatures (50 or more legislators). This is important from a scientific standpoint because it means that the basic geometric model of choice that underlies all the estimation methods produces meaningful results.

The major probabilistic models of parliamentary voting are based on the *random utility model* developed by McFadden (1976). In the random utility model, a legislator's overall utility for voting Yea is the sum of a deterministic utility and a random error. The same is true for the utility of voting Nay. To recap the discussion in Chapter 3, suppose there are p legislators, q roll calls, and s dimensions indexed by i=1,..,p, j=1,...,q, and k=1,…,s, respectively. Legislator i's utility for the Yea outcome on roll call j is:

$$U_{ijy} = u_{ijy} + \varepsilon_{ijy} \qquad (4.1)$$

where u_{ijy} is the *deterministic* portion of the utility function and ε_{ijy} is the *stochastic* or *random* portion of the utility function. Recall from Chapter 3 that if there is no error, then the legislator votes Yea if $U_{ijy} > U_{ijn}$. Equivalently, if the difference, $U_{ijy} - U_{ijn}$, is positive, the legislator votes Yea. With random error the utility difference is:

$$U_{ijy} - U_{ijn} = u_{ijy} - u_{ijn} + \varepsilon_{ijy} - \varepsilon_{ijn}$$

So that the legislator votes Yea if:

$$u_{ijy} - u_{ijn} > \varepsilon_{ijn} - \varepsilon_{ijy}$$

That is, the legislator votes Yea if the difference in the deterministic utilities is greater than the difference between the two random errors. Because the errors are unobserved, we must make an assumption about the error distribution from which they are drawn. Armed with that assumption, we can calculate the *probability* that the legislator will vote Yea. That is,

$$\mathbf{P}(\text{Legislator i Votes Yea}) = \mathbf{P}(U_{ijy} - U_{ijn} > 0) = \mathbf{P}(\varepsilon_{ijn} - \varepsilon_{ijy} < u_{ijy} - u_{ijn})$$

$$\mathbf{P}(\text{Legislator i Votes Nay}) = \mathbf{P}(U_{ijy} - U_{ijn} < 0) = \mathbf{P}(\varepsilon_{ijn} - \varepsilon_{ijy} > u_{ijy} - u_{ijn})$$

so that $\mathbf{P}(\text{Yea})+\mathbf{P}(\text{Nay}) = 1$.[1]

In the next two sections I discuss the functional form of the deterministic portion of the utility function followed by the stochastic portion. These two sections are followed by a discussion of the estimation of the NOMINATE (*Nomina*l *T*hree-Step *E*stimation) (Poole and Rosenthal, 1985; 1991; 1997) and Quadratic-Normal (Poole, 2001) models. This leads into a discussion of the thorny statistical issues alluded to above, namely, parameter proliferation and the consistency of the estimated parameters, and the problem of obtaining the standard errors of the parameters.

The Deterministic Portion of the Utility Function

The two most common deterministic utility functions used in applied work are the normal and the quadratic. The classic normal "bell curve" has the advantage that the utility is always positive, there is decreasing marginal loss, and the utility asymptotes to a fixed value (usually zero). In contrast, the quadratic function has an increasing marginal loss and utility does become infinitely negative as the distance becomes large (see Figure 2.1).

Howard Rosenthal and I based our NOMINATE model on the normal distribution utility function. We decided to use the normal distribution rather than the quadratic because we felt that the normal distribution is a more realistic model of choice behavior.[2] The normal distribution concentrates the utility near the individual's ideal point, with tails that quickly approach zero as the choices become more and more distant. Choices out in the tails beyond two standard deviations all have near-zero utility.

Normally distributed utility is therefore an excellent model for the phenomena of alienation and indifference. Alienation occurs when a person is faced with a set of choices that are all very distant from her ideal point and on the same side of the space.

For example, if she is a moderate liberal and faced with two choices, both of which have conservative outcomes well to her right, she may abstain because of alienation. Similarly, extremist political groups see big differences between themselves and regard all mainstream political parties as "capitalist tools" or "socialist fools," depending on which end of the spectrum they are looking inward from.

Indifference occurs when the individual is faced with a set of choices that "bracket" her in the space. For example, if she is a moderate conservative and she is faced with two choices, one with a very conservative outcome and another with a very liberal outcome, her utility from either outcome is very low because they are under opposite tails of her utility function, and so she may abstain.

In the U.S. Congress most legislators place a high premium on *not* abstaining, because avoiding a vote is often used against them in the next election. Consequently, a legislator's choice in a setting where she is indifferent because of either alienation or indifference may be driven by the stochastic portion of the utility function because the deterministic utility is so tiny. When better multi-choice spatial models of legislative voting are implemented and applied to legislatures where there is policy-related abstention, I expect that the normal utility model will show an unambiguous advantage over the quadratic.

I will discuss the normal distribution utility function first, then the quadratic, and then how the two differ in the estimation of the roll call outcomes.

With the normal distribution utility model, legislator i's utility for the Yea outcome on roll call j is:

$$u_{ijy} = \beta e^{\left(-\frac{1}{2}\sum_{k=1}^{s} w_k d_{ijky}^2\right)} \qquad (4.2)$$

where d_{ijky}^2 is the squared distance of the i^{th} legislator to the Yea outcome on the k^{th} dimension;

$$d_{ijky}^2 = (X_{ik} - O_{jky})^2$$

the w_k are salience weights ($w_k > 0$) ; and because there is no natural metric, β "adjusts" for the overall noise level and is proportional to the variance of the error distribution. The w_k allow the indifference curves of the utility function to be ellipses rather than circles.

The difference between the deterministic utilities is:

$$u_{ijy} - u_{ijn} = \beta \left\{ e^{\left(-\frac{1}{2}\sum_{k=1}^{s} w_k d_{ijky}^2\right)} - e^{\left(-\frac{1}{2}\sum_{k=1}^{s} w_k d_{ijkn}^2\right)} \right\} \qquad (4.3)$$

This rather ungainly equation does not lend itself to any further simplification. But, although it looks complex, computationally it is not difficult to work with.

With the quadratic distribution utility model, legislator i's utility for the Yea outcome on roll call j is just:

$$u_{ijy} = -d_{ijy}^2 = -\sum_{k=1}^{s} \left(X_{ik} - O_{jky}\right)^2 \qquad (4.4)$$

Unlike the normal distribution utility function, the quadratic utility function simplifies nicely. The difference between the deterministic quadratic utilities is:

$$u_{ijy} - u_{ijn} = -\sum_{k=1}^{s} (X_{ik} - O_{jky})^2 + \sum_{k=1}^{s} (X_{ik} - O_{jkn})^2$$

$$= -\sum_{k=1}^{s} X_{ik}^2 + 2\sum_{k=1}^{s} X_{ik}O_{jky} - \sum_{k=1}^{s} O_{jky}^2 + \sum_{k=1}^{s} X_{ik}^2 - 2\sum_{k=1}^{s} X_{ik}O_{jkn} + \sum_{k=1}^{s} O_{jkn}^2$$

$$= -2\sum_{k=1}^{s} X_{ik}(O_{jkn} - O_{jky}) + \sum_{k=1}^{s} (O_{jkn} - O_{jky})(O_{jkn} + O_{jky}) \qquad (4.5)$$

The two terms in equation (4.5) are a sum of cross-products and a sum of squared outcome locations. The cross-product term of equation (4.5) is the foundation for the *directional theory of voting* developed by George Rabinowitz. By separating the two terms in a multivariate framework, it is possible to nest both a directional theory of choice and a *proximity* (distance-based) theory of choice (i.e., the traditional spatial model) in a single framework. Namely, in a regression-like analysis separate coefficients can be estimated for the two terms (Rabinowitz and Macdonald, 1989; Platt, Poole, and Rosenthal 1992; Merrill and Grofman, 1999).[3] If the two coefficients are equal, a proximity model of choice is appropriate. If the coefficient on the cross-product term is large relative to the coefficient on the squared term, it supports the directional model of choice. Platt, Poole, and Rosenthal (1992) applied this model to roll call voting in the 92nd to the 99th U.S. Houses and found that the directional theory can be unambiguously rejected in favor of the traditional spatial model. When we use both a quadratic utility framework with the terms separated as shown above, and the normal utility framework where the squared distances in the exponent of the utility function are broken into a sum of cross-products and a sum of squares, the resulting coefficients are almost exactly equal in all cases.[4]

Returning to equation (4.5), note that the s by 1 vector:

$$\mathbf{O_{jn}} - \mathbf{O_{jy}} = \begin{bmatrix} O_{j1n} - O_{j1y} \\ O_{j2n} - O_{j2y} \\ \vdots \\ O_{jsn} - O_{jsy} \end{bmatrix}$$

is equal to a constant times the normal vector, $\mathbf{N_j}$ because the line joining the outcome points $\mathbf{O_{jn}}$ and $\mathbf{O_{jy}}$ is parallel to the normal vector. Therefore, subtracting $\mathbf{O_{jy}}$ from both points produces a vector that is simply a constant times the normal vector (see Figure 4.1):

$$\gamma_j \mathbf{N_j} = \mathbf{O_{jn}} - \mathbf{O_{jy}} \tag{4.6}$$

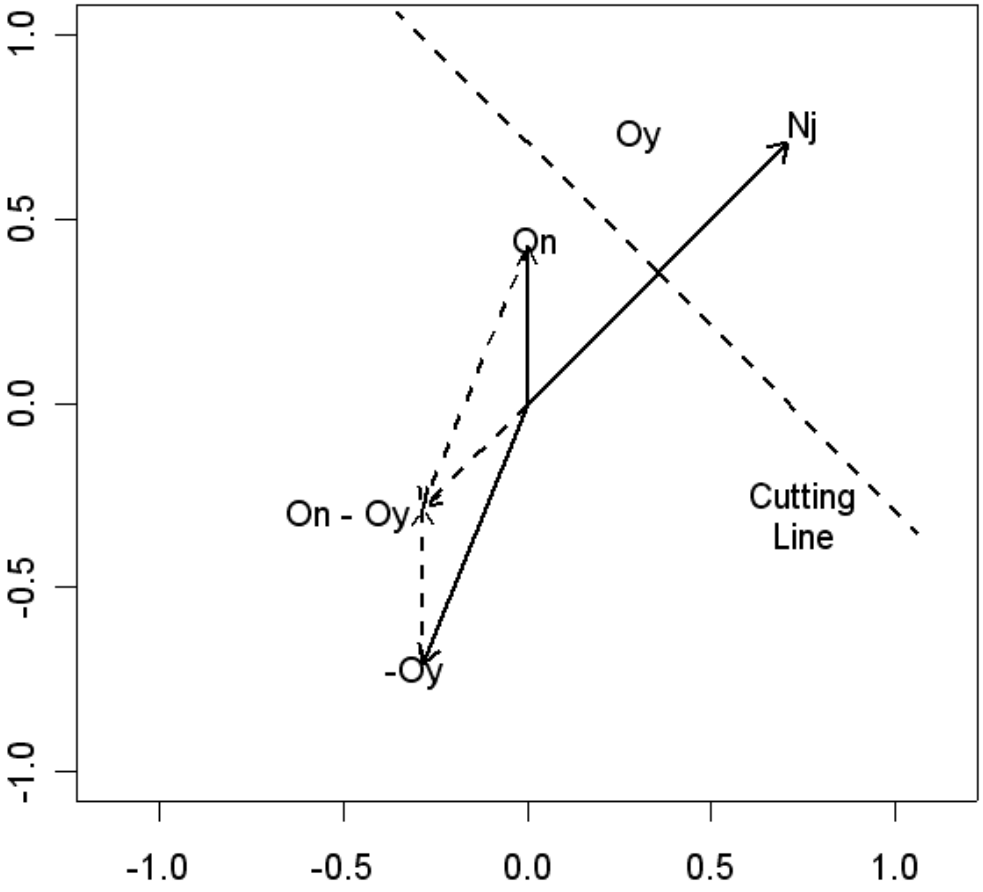
where

$$\gamma_j = \left[\sum_{k=1}^{s} (O_{jkn} - O_{jky})^2 \right]^{\frac{1}{2}} \text{ if } \mathbf{O_{jn}'N_j} > \mathbf{O_{jy}'N_j} \text{ or}$$

$$\gamma_j = -\left[\sum_{k=1}^{s} (O_{jkn} - O_{jky})^2 \right]^{\frac{1}{2}} \text{ if } \mathbf{O_{jn}'N_j} < \mathbf{O_{jy}'N_j}$$

γ_j is the *directional distance* between the Yea and Nay outcomes in the space.[5]

Figure 4.1: Vector Difference of Outcome Coordinates



The s by 1 vector

$$\mathbf{O_{jn}} + \mathbf{O_{jy}} = \begin{bmatrix} O_{j1n} + O_{j1y} \\ O_{j2n} + O_{j2y} \\ \vdots \\ O_{jsn} + O_{jsy} \end{bmatrix}$$

divided by 2 is simply the s by 1 vector of midpoints for the Yea and Nay outcomes for roll call j. That is:

$$\mathbf{Z_j} = \frac{\mathbf{O_{jn}} + \mathbf{O_{jy}}}{2}$$

This allows equation (4.5) to be rewritten as the vector equation:

$$u_{ijy} - u_{ijn} = 2\gamma_j(\mathbf{Z_j'N_j} - \mathbf{X_i'N_j}) = 2\gamma_j(c_j - w_i) \qquad (4.7)$$

where w_i is the projection of the i^{th} legislator's ideal point onto the line defined by the normal vector, $\mathbf{N_j}$ and its reflection $-\mathbf{N_j}$ (see Figure 3.8), and c_j is the projection of the midpoint of the roll call outcomes onto the line defined by the normal vector, $\mathbf{N_j}$ and its reflection $-\mathbf{N_j}$. If the legislator votes Yea and she is on the same side of the cutting plane as the Yea outcome, then from Equation (4.7):

if $\gamma_j > 0$ and $w_i < c_j$, or

if $\gamma_j < 0$ and $w_i > c_j$, then $u_{ijy} > u_{ijn}$

In one dimension, $\mathbf{N_j}$ can be set equal to 1 and $\gamma_j = O_{jn} - O_{jy}$. Hence equation (4.7) becomes simply $2(O_{jn} - O_{jy})(c_j - x_i) = 2\gamma_j(x_i - c_j)$.[6]

In more than one dimension only the cutting line and the distance between the two outcome points are identified in the quadratic utility model; that is, the directional distance γ_j, the normal vector $\mathbf{N_j}$, and the projection of the midpoint $\mathbf{Z_j}$ on the normal vector, c_j. Although it appears that the number of parameters for a roll call is s+2, given the fact that the normal vector has unit length, that is, $\mathbf{N_j'N_j}=1$, the normal vector is completely determined by the s-1 angles of the vector from the coordinate axes so that the actual number of parameters is s+1. In a polar coordinate system the s-1 angles produce s coordinates provided that the vector is of fixed length. For example, in two dimensions the normal vector can always be written as $\begin{bmatrix}\cos\theta \\ \sin\theta\end{bmatrix}$ where θ is the angle of the vector from the horizontal axis and $0 \le \theta \le 2\pi$; in three dimensions the normal vector

can always be written as $\begin{bmatrix} \sin\varphi\cos\theta \\ \sin\varphi\sin\theta \\ \cos\varphi \end{bmatrix}$ where θ is the angle in the plane of the first and second dimensions, $0 \leq \theta \leq 2\pi$, and φ is the angle from the third dimension to the plane, $0 \leq \varphi \leq \pi$.

In sum, in the quadratic utility model each roll call is determined by s+1 parameters – γ_j, $\mathbf{c_j}$, and the s-1 angles. As a practical problem, however, it is easier simply to estimate the normal vector, $\mathbf{N_j}$, directly rather than parameterize the problem in terms of the underlying angles, because techniques similar to the cutting plane procedure in Chapter 3 are easy to implement in probabilistic voting problems.

In more than one dimension the fact that each roll call is determined by only s+1 parameters means that the outcome coordinates are identified only up to *parallel tracks through the space*. Figure 4.2 shows the geometry.

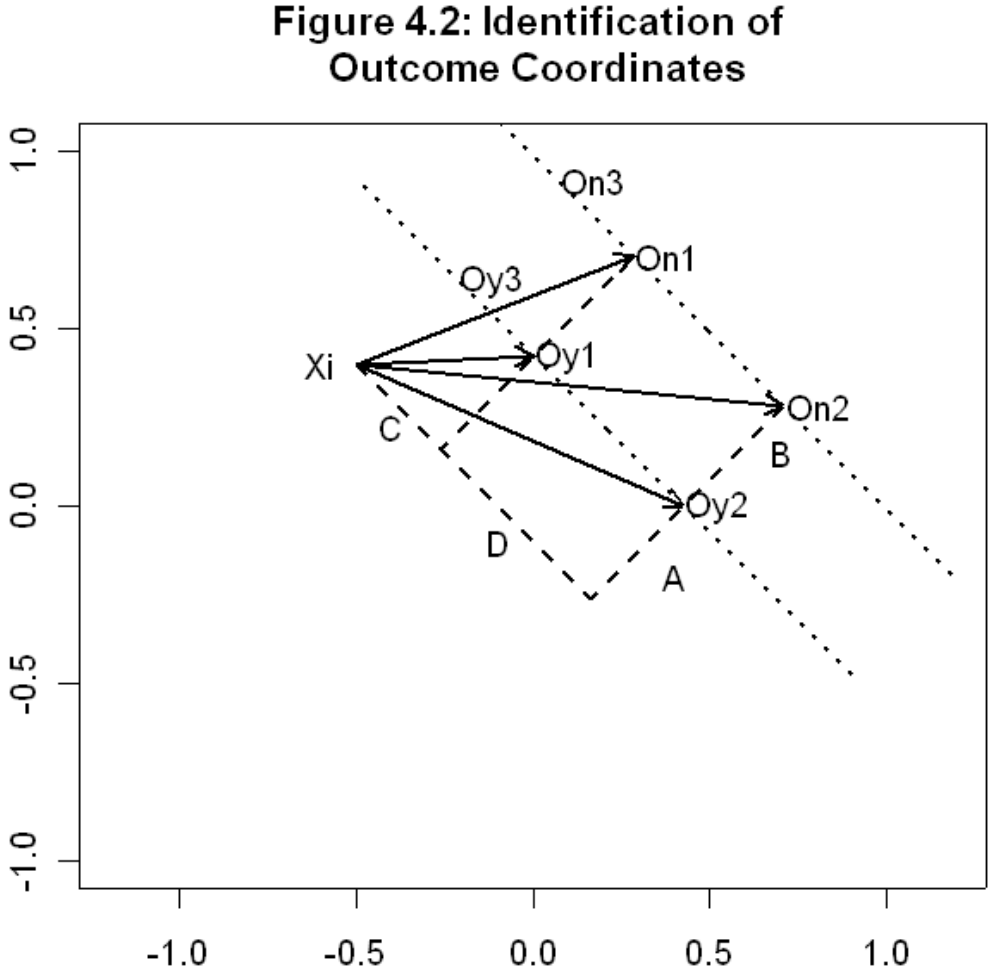


Figure 4.2 shows one legislator, $\mathbf{X_i}$, and her distances to two pairs of outcomes (the vectors in the figure). The parallel dotted lines pass through the two pairs of outcomes. By the Pythagorean Theorem, the squared distance between $\mathbf{X_i}$ and $\mathbf{O_{y1}}$ is $C^2 + A^2$ (see the indicated line segments in the figure). Similarly, the squared distance between $\mathbf{X_i}$ and $\mathbf{O_{n1}}$ is $C^2 + (A+B)^2$. Therefore,

$$u_{iy} - u_{in} = -d_{iy}^2 + d_{in}^2 = -(C^2 + A^2) + [C^2 + (A+B)^2] = 2AB + B^2$$

Similarly, the squared distance between $\mathbf{X_i}$ and $\mathbf{O_{y2}}$ is $(C+D)^2 + A^2$ and the squared distance between $\mathbf{X_i}$ and $\mathbf{O_{n2}}$ is $(C+D)^2 + (A+B)^2$. Therefore,

$$u_{iy} - u_{in} = -d_{iy}^2 + d_{in}^2 = -[(C+D)^2 + A^2] + [(C+D)^2 + (A+B)^2] = 2AB + B^2$$

In sum, in more than one dimension the outcome coordinates in the quadratic utility model are identified only up to the parallel tracks shown in Figure 4.2. The disadvantage is that the outcome points can never be definitively estimated. The advantage is the simplicity of the geometry. With the exception of the signed distance parameter, γ_j, the geometry of the quadratic utility function is identical to that shown in Chapter 2, and the OC method shown in Chapter 3 can be used as a basis for estimating the probabilistic model.

In contrast, in the normal distribution utility function the outcome coordinates are identified, but the price is 2s parameters – the two outcome points $\mathbf{O_{jn}}$ and $\mathbf{O_{jy}}$. Holding the distance between the two outcomes fixed as in Figure 4.2, $u_{iy} - u_{in}$ will vary as the outcomes are moved up and down the tracks, and it is a maximum when the outcomes are located at $\mathbf{O_{y3}}$ and $\mathbf{O_{n3}}$. This identification is due to the non-linearity of the utility difference. For any legislator with the distance between the two outcomes held fixed, the utility difference is maximized when the legislator's ideal point and the outcomes lie on a line parallel to the normal vector for the roll call cutting plane. In the example shown in Figure 4.2, the utility difference equation (4.3) for the $\mathbf{O_{y3}}$ and $\mathbf{O_{n3}}$ outcome points is:

$$u_{iy} - u_{in} = e^{-A^2} - e^{-(A+B)^2} = e^{-A^2}\left(1 - e^{-(2AB+B^2)}\right)$$

The utility difference for the $\mathbf{O_{y1}}$ and $\mathbf{O_{n1}}$ outcome points is:

$$u_{iy} - u_{in} = e^{-(A^2+C^2)} - e^{-[C^2+(A+B)^2]} = e^{-(A^2+C^2)}\left(1 - e^{-(2AB+B^2)}\right)$$

Now, because $e^{-A^2} > e^{-(A^2+C^2)}$ the legislator's utility difference is always maximized when her ideal point and the outcomes lie on a line parallel to the normal vector.

Not every legislator can be on the line running through the outcome points! Consequently, *if we hold the distance between the outcomes fixed*, the outcomes will be positioned on the tracks so that the total of the utility differences of all the legislators vis a vis their chosen outcomes is maximized. In contrast, the total of the utility differences of all the legislators in the quadratic model is the same regardless of the location of the outcome points on the tracks.

Another crucial difference between the two utility functions is how they are affected by an increase in the distance between the two outcomes. Note that as the distance, B, in Figure 4.2 increases, the quadratic utility difference for $\mathbf{X_i}$ increases monotonically. That is, if $B \to \infty$ then $A \to \infty$ and $2AB + B^2 \to \infty$. However, for the normal utility difference if $B \to \infty$ then $A \to \infty$ and $e^{-A^2} \to 0$ and $e^{-(A^2+B^2)} \to 0$ so that the utility difference goes to zero. This means that for the normal utility model the distance between the two outcomes cannot "explode" – that is, there is a value beyond which the utility difference (and the probability of the choice) begins to decline. This fact is illustrated in Figure 4.3.

Figure 4.3: Utility Difference Normal Model vs. Distance Between Outcomes

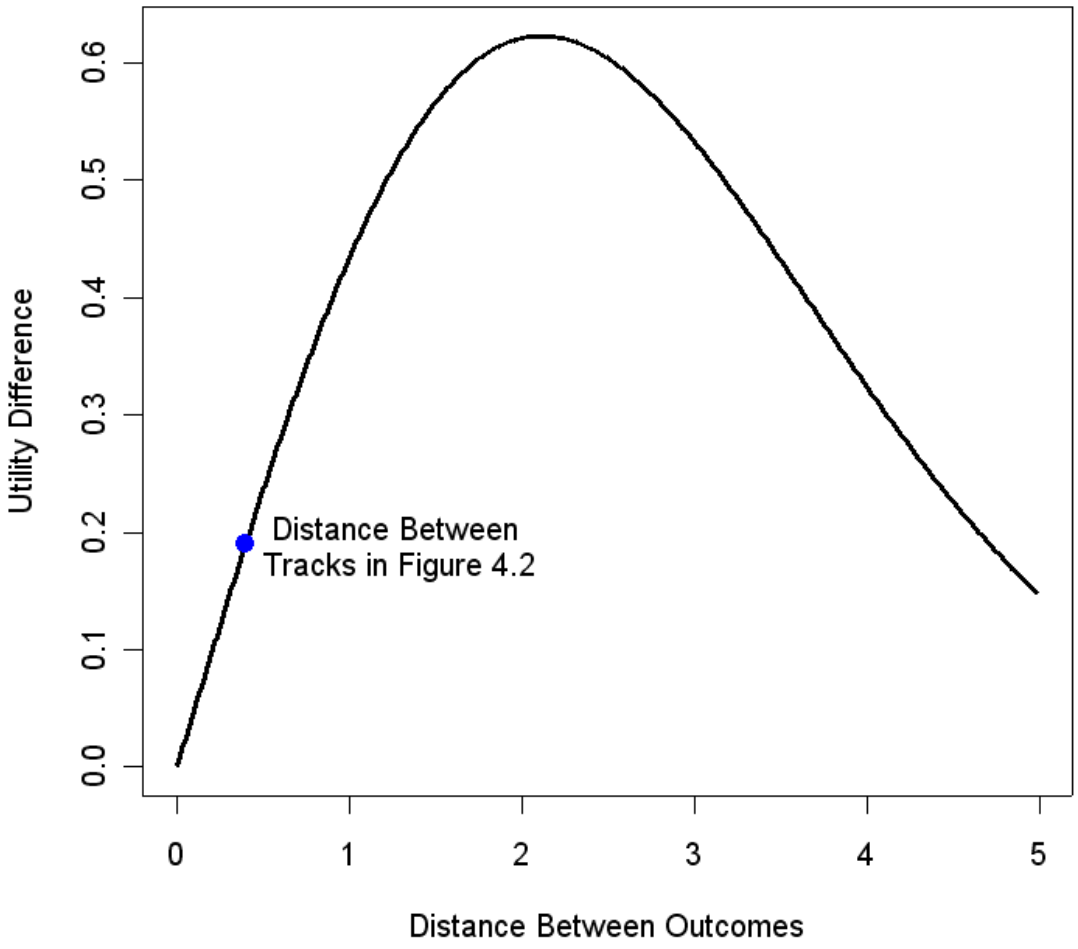


Figure 4.3 shows the effect of increasing the distance between $\mathbf{O}_{y3}$ and $\mathbf{O}_{n3}$ in Figure 4.2 from 0 to 5 with legislator $\mathbf{X}_i$ held fixed. The distance between the "tracks" in Figure 4.2 is .4 and the corresponding utility difference is about .19. The utility difference peaks at a distance of 2.12 with a value of .623 and then declines thereafter.

The Stochastic Portion of the Utility Function

When we turn to the stochastic portion of the utility function stated in equation (4.1) above, three probability distributions have been used to model the error; the normal

(Ladha, 1991; McCarty, Poole, and Rosenthal, 1997; Londregan, 2000; Jackman, 2000; Poole, 2001), uniform (Heckman and Snyder, 1997), and logit (Poole and Rosenthal, 1985; 1991; 1997). The normal is clearly the best from both a theoretical *and* a behavioral standpoint.

From a statistical standpoint, the random errors associated with the Yea and Nay choices should be a *random sample* (i.e., a set of independent and identically distributed random variables) from a known distribution. If ε_{ijn} and ε_{ijy} are a random sample from a known distribution, then the probability distribution of the difference between the two random errors, $\varepsilon_{ijn} - \varepsilon_{ijy}$, is usually easy to derive. From a behavioral standpoint, it seems sensible to assume that the distributions of ε_{ijn} and ε_{ijy} are *symmetric and unimodal* and that ε_{ijn} and ε_{ijy} are uncorrelated. The normal distribution is the only one of the three distributions to satisfy all these criteria. To illustrate, assume that ε_{ijn} and ε_{ijy} are drawn (a random sample of size two) from a normal distribution with mean zero and variance one-half. The difference between the two errors has a standard normal distribution; that is

$$\varepsilon_{ijn} - \varepsilon_{ijy} \sim N(0, 1)$$

and the distribution of the difference between the overall utilities is

$$U_{ijy} - U_{ijn} \sim N(u_{ijy} - u_{ijn}, 1)$$

Hence the probability that legislator i votes Yea on the j^{th} roll call can be rewritten as:

$$\mathbf{P_{ijy} = P(}U_{ijy} > U_{ijn}\mathbf{) = P(}\varepsilon_{ijn} - \varepsilon_{ijy} < u_{ijy} - u_{ijn}\mathbf{) = \Phi[}u_{ijy} - u_{ijn}\mathbf{]} \qquad (4.8)$$

Heckman and Snyder (1997) assume that $\varepsilon_{ijn} - \varepsilon_{ijy}$ has a uniform distribution. This is an extremely problematic assumption because ε_{ijn} and ε_{ijy} *cannot be a random sample*! Heckman and Snyder (1997) acknowledge that no distribution exists such that the

probability distribution of the difference between two random draws has a uniform distribution. For example, if ε_{ijn} and ε_{ijy} are drawn from a uniform distribution, then the distribution of their difference will be a triangle-shaped distribution. Assuming that ε_{ijn}-ε_{ijy} has a uniform distribution enables Heckman and Snyder to develop a linear probability model, but the price for this simplicity is that they have no intuitive basis for a behavioral model.

Poole and Rosenthal (1985, 1991, 1997) assume that ε_{ijn} and ε_{ijy} are a random sample from the log of the inverse exponential distribution[7]; that is:

$$f(\varepsilon) = e^{-\varepsilon}e^{-e^{-\varepsilon}} \text{ where } -\infty < \varepsilon < +\infty$$

Dhrymes (1978, pp. 340-352) shows that the distribution of ε_{ijn} - ε_{ijy} is:

$$f(z) = \frac{e^{-z}}{(1 + e^{-z})^2} \text{ where } z = \varepsilon_{ijn} - \varepsilon_{ijy} \text{ and } -\infty < z < +\infty$$

This is the logit distribution.

The log of the inverse exponential distribution and the logit distribution are unimodal *but not symmetric*. Neither distribution is terribly skewed, however, and the distribution function of the logit distribution is reasonably close to the normal distribution function. Namely, integrating f(z) from $-\infty$ to $u_{ijy} - u_{ijn}$ yields the probability that the legislator votes Yea:

$$P_{ijy} = P(U_{ijy} > U_{ijn}) = P(\varepsilon_{ijn} - \varepsilon_{ijy} < u_{ijy} - u_{ijn}) = \int_{-\infty}^{u_{ijn} - u_{ijy}} \frac{e^{-z}}{(1+e^{-z})^2} dz = \frac{e^{u_{ijy}}}{e^{u_{ijy}} + e^{u_{ijn}}} \qquad (4.9)$$

The probabilistic models summarized by equations (4.8) and (4.9) implicitly assume that the underlying error variance is homoskedastic – that is, it is constant across

legislators and across roll calls. A more realistic assumption is that the error variance varies across the roll call votes and across the legislators. In reality, however, for the roll calls, it is impossible to distinguish between the underlying unknown error variance and the distance between the Yea and Nay alternatives (Ladha, 1991; Poole and Rosenthal, 1997; Londregan, 2000). The intuition behind this statement is straightforward. As the distance between the Yea and Nay alternatives increases, it becomes easier for legislators to distinguish between the two policy outcomes, and so they are much less likely to make a voting error. Conversely, if the Yea and Nay alternatives are very close together, then the utility difference is small and it is more likely that voting errors occur.

Increasing/decreasing the distance is equivalent to decreasing/increasing the variance of the underlying error. For example, for the quadratic utility model with normally distributed error, the probability of the observed choice is:

$$\Phi[2\gamma_j(c_j - w_i)] \qquad (4.10)$$

where $2\gamma_j(c_j - w_i) > 0$ if the legislator is on the same side of the cutting plane as the policy outcome she voted for (this is picked up by the signed distance term, γ_j). If all the legislators are correctly classified, then $|\gamma_j| \to \infty$ and $\Phi[2\gamma_j(c_j - w_i)] \to 1$ for all legislators. This situation is equivalent to setting the signed distance term to a constant and writing the probability of the observed choice in the more traditional way:

$$\Phi\left[\frac{2\gamma_j(c_j - w_i)}{\sigma_j}\right]$$

where σ_j^2 is the roll call specific variance, that is:

$$\varepsilon_{ijn} - \varepsilon_{ijy} \sim N(0, \sigma_j^2)$$

Clearly, as $\sigma_j \to 0$ the probabilities on the j^{th} roll call all go to one.

For the normal utility model the probability of voting Yea is:

$$\Phi\left[\beta\left\{e^{\left(-\frac{1}{2}\sum_{k=1}^{s} w_k d_{ijky}^2\right)} - e^{\left(-\frac{1}{2}\sum_{k=1}^{s} w_k d_{ijkn}^2\right)}\right\}\right] \qquad (4.11)$$

Here the interaction of the distance between the two outcomes and the error variance for the roll call is not as clear cut, but it is still there. As I explained above, if the roll call cutting plane perfectly classifies the legislators, then in the normal utility model the distance between the two outcome points will grow large – typically both outcomes are outside the space spanned by the legislators – but will not explode. Note that if the overall noise parameter, β, was allowed to vary across roll calls -- that is, if there were q β_j's -- then these β_j's would be roll call specific variances; that is, $\beta_j = 1/\sigma_j$. Howard Rosenthal and I experimented with this model in some depth in the 1980s. We concluded that even though the distance between the outcome coordinates did not explode, the interaction between the β_j's and the distances was such that estimates of the β_j's were not reliable.

In sum, the roll call specific error variances are picked up by the distance between the two outcome points in both the quadratic and normal utility models.

Legislator specific variance is another matter. In principle, it can be disentangled. For the quadratic utility model the probability of the observed choice is:

$$\Phi\left[\frac{2\gamma_j(c_j - w_i)}{\sigma_i}\right] \qquad (4.12)$$

where σ_i^2 is the legislator specific variance, that is:

$$\varepsilon_{ijn} - \varepsilon_{ijy} \sim N(0,\ \sigma_i^2)$$

Because each γ_j is estimated with respect to the p legislators and each σ_i is estimated with respect to the q roll calls, in principle the two sets of parameters do not have an interaction problem like that between γ_j and σ_j. But because they enter equation (4.12) as a *ratio*, the γ_j's and σ_i's are identified only up to a multiplicative positive constant. That is, for any k> 0, $\frac{2\gamma_j(c_j-w_i)}{\sigma_i} = \frac{2k\gamma_j(c_j-w_i)}{k\sigma_i}$. In practice this identification problem is not a serious concern, because we are primarily interested in how noisy legislators are *relative* to one another. The model shown in equation (4.12) is the basis of the *Quadratic-Normal* (QN) scaling program (Poole, 2001). I will discuss it in more detail later in this chapter.

Legislator-specific variance can also be estimated in the normal utility model. The probability of the observed choice is:

$$\Phi\left[\frac{e^{\left(-\frac{1}{2}\sum_{k=1}^{s} w_k d_{ijky}^2\right)} - e^{\left(-\frac{1}{2}\sum_{k=1}^{s} w_k d_{ijkn}^2\right)}}{\sigma_i}\right] \qquad (4.13)$$

In this formulation the overall noise parameter, β, is unnecessary because its effect will be picked up by the p σ_i's. This model has never been developed in computer code, and it awaits an adventurous researcher![8]

Estimation of Probabilistic Spatial Voting Models

In classical statistical inference we assume that we know the type of distribution (e.g, Normal, Poisson, Bernoulli, etc.) from which our random sample is drawn, and on the basis of that random sample we must infer the values of the *parameters* of the

distribution. This is parametric estimation. Within this classical framework, I assume that the distribution of the difference between the utility for Yea and the utility for Nay for the i^{th} legislator on the j^{th} roll call is normally distributed with constant variance; that is:

$$U_{ijy} - U_{ijn} \sim N(u_{ijy} - u_{ijn}, \sigma^2) \qquad (4.14)$$

The parameters to be estimated on the basis of a random sample from this distribution are the ps legislator coordinates, X_{ik}, the 2qs roll call outcome parameters, O_{jky} and O_{jkn}, the utility function parameters, β and/or the s w_k. The approach I take is classical maximum likelihood; that is, I attempt to find parameter values that maximize the likelihood of actually observing the roll call data. Technically, the likelihood function is:

$$L = \prod_{i=1}^{p} \prod_{j=1}^{q} \prod_{\tau=1}^{2} P_{ij\tau}^{C_{ij\tau}} \qquad (4.15)$$

where τ is the index for Yea and Nay, $P_{ij\tau}$ is the probability of voting for choice τ as given by equations (4.10) to (4.13), and $C_{ij\tau} = 1$ if the legislator's actual choice is τ and zero otherwise. It is standard practice to work with the natural log of the likelihood function:

$$\mathcal{L} = \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{\tau=1}^{2} C_{ij\tau} \ln P_{ij\tau} \qquad (4.16)$$

In the conventional approach, the maximum of equation (4.16) is found by taking the first derivatives of the log of the likelihood function with respect to all the parameters, setting the first derivatives equal to zero, and then solving the set of equations for the parameter values. The matrix of second derivatives is also computed to ensure that the solution is indeed a maximum.

This conventional approach does not work well for two reasons: the number of parameters that must be estimated and the necessity for constraints on the parameter values.

Even for relatively small legislatures, the number of parameters that must be estimated is very large. For example, a typical recent U.S. Senate has 100 Senators casting 500 or more roll call votes. In two dimensions, about 2,200 parameters would have to be estimated simultaneously. Using standard maximization methods would require repeated inversions of matrices of 2,200 by 2,200. Even with modern computers this is a formidable undertaking, and there is good reason to believe that the approach would not work very well in any event. For one thing, how the level of error *interacts* with the underlying geometry is not yet clearly understood (see Chapter 7). Without knowing this relationship we cannot be certain about the interval scale information that is being recovered by the scaling program and how this affects the matrix of second derivatives. There may be hidden linearities that we do not yet understand. Empirically, a matrix of this size it is very likely to be singular because of "accidental" linear combinations of rows/columns.

The second reason the conventional approach does not work is that constraints must be placed on the legislators and roll calls. For example, if the error level is low and some extreme legislators *always* vote for the "liberal"/"conservative" outcome *on every roll call*, their ideal point cannot be identified beyond the fact that they are *at the edge of the space*. A similar problem crops up for perfect roll calls and for lopsided noisy roll calls. In all three cases the parameters must not be allowed to take on absurd values so that the estimated parameters are *not* at the maximum of the log-likelihood function.

The approach that Howard Rosenthal and I developed that we dubbed NOMINATE – *Nomina*l *T*hree-Step *E*stimation -- solves the estimation problem by dividing the parameters into three sets and estimating each set separately with the other two held fixed. Given reasonable starting values for the legislator ideal points, we can estimate the roll call parameters *given* these ideal points. Similarly, *given* estimates of the roll call parameters, we can obtain better estimates of the ideal points. The third set is the utility function parameters that are estimated *given* both the legislator ideal points and the roll call outcome points. NOMINATE cycles through the three sets until they reproduce each other.[9] That is, at convergence each set is reproduced by the other two -- the parameters no longer change.

The Quadratic-Normal (QN) scaling method uses a quadratic deterministic utility function and normally distributed error. It is based on OC and it also solves the estimation problem by dividing the parameters into three sets and cycling through them until the three reproduce each other. The parameters are identical to OC with the addition of the signed distances and legislative specific variances.

The NOMINATE (Normal-Normal) Model

In our book Howard Rosenthal and I (1997) detail the original D-NOMINATE and W-NOMINATE algorithms that assumed that error followed the logit model shown in equation (4.9). When we began our work in late 1982, we used the logit model because of computer speed and memory limitations. Calculating logit probabilities was easier and faster because *the logit probability is given by a formula*. In contrast, in the

older computers normal probabilities were calculated by a machine subroutine that used a power series approximation (see Chapter 5).[10]

When NSF announced the supercomputer initiative in the summer of 1985, we were granted time on the Cyber 205 vector supercomputer at Purdue University to develop NOMINATE further. The architecture of the Cyber 205 – its use of vector pipelines (now in the all the latest Intel CPUs) -- and the programming language – VECTOR FORTRAN – that was implemented on it were ideally suited to large discrete choice problems.[11] Consequently, we were able to estimate *dynamic* multidimensional spatial models, and we developed D-NOMINATE over the 1986-1988 period.

The original NOMINATE was a one-dimensional program, and it was re-written by Nolan McCarty in 1991-92 to analyze more than one dimension. This version of NOMINATE became known as W-NOMINATE. The "W" stands for "weighted" because it uses the weighted distance model shown in equation (4.3). The weights allowed for elliptical indifference curves for the utility function rather than circles.

By the mid-1990s personal computers became powerful enough in processing speed and memory that it became feasible to implement a dynamic multidimensional spatial model on personal computers. DW-NOMINATE is based on normally distributed errors rather than logit errors, and it uses the W-NOMINATE weighted distance model (McCarty, Poole, and Rosenthal, 1997). DW-NOMINATE is structured to analyze any number of Congresses and can be applied to a single Congress. Consequently, I will focus on DW-NOMINATE in my discussion below rather than the older algorithms.

Although DW-NOMINATE, like D-NOMINATE, was designed to deal with the U. S. House and Senate, it can be used to analyze any voting body that meets in multiple

sessions over time; for example, U. S. state legislatures, the United Nations, the European Parliament, and so on. Below I use the term "Parliamentary Session" to refer to any distinct session of a voting body. In the U.S. context this would be one "Congress" or a two year period.

Let T be the number of Parliamentary Sessions that are indexed by t=1,...,T; s denote the number of policy dimensions (k=1,...,s); p_t denote the number of legislators in Parliamentary Session t (i=1,...,p_t); q_t denote the number of roll call votes in Parliamentary Session t (j=1,...,q_t); and T_i denote the number of Parliamentary Sessions in which legislator i served (t=1,...,T_i). To allow for spatial movement over time, the legislator ideal points are treated as being polynomial functions of time; namely, legislator i's coordinate on dimension k at time t is given by:

$$X_{ikt} = \chi_{ik0} + \chi_{ik1}\mathcal{T}_{t1} + \chi_{ik2}\mathcal{T}_{t2} + \ldots + \chi_{ikv}\mathcal{T}_{tv} \qquad (4.17)$$

where v is the degree of the polynomial, the χ's are the coefficients of the polynomial, and the time-specific terms -- the $\mathcal{T}$s -- are Legendre polynomials.[12] Specifically, the first three terms of a Legendre polynomial representation of time are:

$$\mathcal{T}_{t1} = -1+(t-1)\frac{2}{(T_i-1)}$$

$$\mathcal{T}_{t2} = \frac{3\mathcal{T}_{t1}^2-1}{2} \qquad t=1,\ldots,T_i \qquad (4.18)$$

$$\mathcal{T}_{t3} = \frac{5\mathcal{T}_{t1}^3-3\mathcal{T}_{t1}}{2}$$

Legendre polynomials have the nice property of being orthogonal on the interval [-1,+1]. This is convenient because DW-NOMINATE scales the legislators and roll call

midpoints to be in the unit hypersphere (more on this below). The orthogonality of the Legendre polynomials is a *continuous* property, but even with discrete data the linear and quadratic terms will be orthogonal. If ordinary powers of time were used instead, that is, t, t^2 , t^3 , and so on, these would be *correlated*. For example, suppose a legislator serves in five Parliamentary Sessions. Then her ideal point in two dimensions using a *cubic* model (v=3) can be written as the matrix product:

$$\begin{bmatrix} 1 & -1 & 1 & -1 \\ 1 & -.5 & -.125 & .4375 \\ 1 & 0 & -.5 & 0 \\ 1 & .5 & -.125 & -.4375 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \chi_{i10} & \chi_{i20} \\ \chi_{i11} & \chi_{i21} \\ \chi_{i12} & \chi_{i22} \\ \chi_{i13} & \chi_{i23} \end{bmatrix} = \begin{bmatrix} X_{i11} & X_{i21} \\ X_{i12} & X_{i22} \\ X_{i13} & X_{i23} \\ X_{i14} & X_{i24} \\ X_{i15} & X_{i25} \end{bmatrix}$$

or

$$\mathcal{T}\boldsymbol{\chi}_{\mathbf{i}} = \mathbf{X}_{\mathbf{iT}} \qquad (4.19)$$

If only one Parliamentary Session is being estimated, the legislator ideal points are simply $\mathbf{X_i} = \begin{bmatrix} X_{i1} \\ X_{i2} \\ . \\ . \\ X_{is} \end{bmatrix}$ as before.

The two roll call outcome points associated with Yea and Nay on the k^{th} dimension at time t can be written in terms of their midpoint and the distance between them; namely,

$$\begin{aligned} O_{jkyt} &= Z_{jkt} - \delta_{jkt} \\ O_{jknt} &= Z_{jkt} + \delta_{jkt} \end{aligned} \qquad (4.20)$$

where the midpoint is:

$$Z_{jkt} = \frac{(O_{jkyt} + O_{jknt})}{2}$$

and δ_{jkt} is *half* the signed distance between Yea and Nay points on the k^{th} dimension (note that δ_{jkt} can be negative); that is

$$\delta_{jkt} = \frac{(O_{jkyt} - O_{jknt})}{2} \qquad (4.21)$$

Recall that by Figure 1 and equation (4.6), the s by 1 vector, $\boldsymbol{\delta_{jt}}$, is equal to a constant times the normal vector for the roll call; that is,

$$\kappa \mathbf{N_j} = \boldsymbol{\delta_{jt}}$$

so that

$$c_j = (1/\kappa)\mathbf{Z_{jt}}'\boldsymbol{\delta_{jt}}$$

This shows that the basic underlying geometry is reflected in Equations (4.20) and (4.21).

The distance of legislator i to the Yea outcome on the k^{th} dimension for roll call j at time t is:

$$d^{2}_{ijkyt} = (X_{ikt} - O_{jkyt})^2$$

Legislator i's utility for the Yea outcome on roll call j at time t is:

$$U_{ijyt} = u_{ijyt} + \varepsilon_{ijyt} = \beta e^{\left(-\frac{1}{2}\sum_{k=1}^{s} w_k^2 d^2_{ijkyt}\right)} + \varepsilon_{ijyt} \qquad (4.22)$$

where u_{ijyt} is the deterministic portion of the utility function, ε_{ijyt} is the stochastic portion, and w_k are the salience weights. Because the stochastic portion of the utility function is normally distributed with constant variance, β is proportional to $1/\sigma^2$ where:

$$\varepsilon \sim N(0, \sigma^2)$$

Hence the probability that legislator i votes Yea on the j^{th} roll call at time t is the same as equation (4.11) except for the time subscript; specifically

$$P_{ijyt} = P(U_{ijyt} > U_{ijnt}) = P(\varepsilon_{ijnt} - \varepsilon_{ijyt} < u_{ijyt} - u_{ijnt}) = \Phi(u_{ijyt} - u_{ijnt}) =$$

$$\Phi\left[\beta\left\{e^{\left(-\frac{1}{2}\sum_{k=1}^{s} w_k d_{ijkyt}^2\right)} - e^{\left(-\frac{1}{2}\sum_{k=1}^{s} w_k d_{ijknt}^2\right)}\right\}\right] \qquad (4.23)$$

The natural log of the likelihood function is:

$$\mathcal{L} = \sum_{t=1}^{T}\sum_{i=1}^{p_t}\sum_{j=1}^{q_t}\sum_{\tau=1}^{2} C_{ij\tau t} \ln P_{ij\tau t} \qquad (4.24)$$

where τ is the index for Yea and Nay, $P_{ij\tau t}$ is the probability of voting for choice τ as given by equation (4.23), and $C_{ij\tau t} = 1$ if the legislator's actual choice is τ and zero otherwise.

The total number of parameters to be estimated is at most:

$$2s\sum_{t=1}^{T} q_t + sp(v+1) + s$$

where $2s\sum_{t=1}^{T} q_t$ is the number of roll call parameters, sp(v+1) is the *maximum possible* number of legislator parameters – that is, all p legislators serve in at least v+1 Parliamentary Sessions so that the v+1 χ_{ik}'s can be estimated -- and the "s" is β and w_2 through w_s (the weight on the first dimension can be set equal to 1). In practice v is usually 0 – the *constant model* where the legislator's ideal point is the same in every Parliamentary Session in which she serves – or v is 1 – the *linear model* in which legislators are allowed to follow a straight line trajectory through the space over time. In our original D-NOMINATE work, Howard Rosenthal and I found that the linear model in

two dimensions was the best combination of explanatory power and the number of parameters (Poole and Rosenthal, 1991, 1997).

Because the roll calls are nominal data – just Yeas, Nays, and not voting – the spatial model outlined above has no natural metric. Some metric must be chosen! The metric used by DW-NOMINATE is to confine the *constant point of the legislator* – that is, the s by 1 vector $\boldsymbol{\chi}_{i0}$ from the s polynomials defined by equation (4.17) -- and the midpoints of the roll call outcomes – the $\mathbf{Z}_j$'s – to be within a unit hypersphere. These constraints are similar to those that Howard Rosenthal and I used in D-NOMINATE (Poole and Rosenthal, 1997, Appendix A).

The unit hypersphere constraint has the nice consequence that the normal vectors – the $\mathbf{N}_j$, which are part of the underlying geometry – have the same length as the radius of the space. This is computationally convenient *but not necessary*. It just makes estimation easier. Note that confining the constant term of the legislator's polynomials to be within the unit hypersphere means that it is possible for extreme legislators with large time trends to go somewhat outside the boundaries. This is not a serious problem in practice, because the roll call midpoints are confined within the unit hypersphere so that the legislator coordinates do not "explode" – that is, legislators who always vote for the "liberal"/"conservative" outcome *on every roll call* do not have ideal points in outer space.

The parameters of the model are estimated in the following three-step process:

Step 0: Generate Starting Estimates for the Legislator Ideal Points, X_{ik} and the Roll Call Parameters -- the Z_{jkt} and δ_{jkt}

Step 1: Estimate the Utility Function Parameters – β, w_2, … , w_s – with the legislator and roll call parameters held fixed

Step 2: Estimate the Roll Call Parameters -- the Z_{jkt} and δ_{jkt} -- with the legislator and utility function parameters held fixed

Step 3: Estimate the Legislator Ideal Points – the χ_{ik0} through χ_{ikv} – with the roll call and utility parameters held fixed

Go to Step 1

I will discuss each step in turn.

Step 0 is in many respects the most important. Because the likelihood function is non-linear, half the battle is in the starting estimates of the parameters. If the number of Parliamentary Sessions is small, the best method of obtaining starting values for the legislator ideal points is the method outlined in Chapter 2. First, compute the p by p agreement score matrix; second, convert the agreement score matrix into a matrix of squared distances; three, double-center the matrix of squared distances; and finally perform an eigenvalue-eigenvector decomposition of the double-centered matrix. The s-dimensional starting values are:

$$\tilde{\mathbf{X}} = \eta \mathbf{U}_s \qquad (4.25)$$

where $\tilde{\mathbf{X}}$ is the p by s matrix of legislator starting coordinates. $\mathbf{U}$ is p by s matrix of eigenvectors from the decomposition of the double-centered matrix such that $\mathbf{U'U}=\mathbf{I_s}$ where $\mathbf{I_s}$ is an s by s identity matrix, and η is a positive constant that is an *inflation factor* that scales the starting coordinates so that the legislator who is farthest from the origin lies on the rim of the unit hypersphere. I will discuss the practical aspects of equation (4.25) in detail in Chapter 5, including how to deal with missing data – that is, two

members who do not overlap so no agreement score can be computed – and how to decide on the number of dimensions to estimate.

These starting estimates for the legislators produce a single ideal point for each legislator; that is, the constant model, v=0, is assumed initially in order to get the overall estimation started. At Step 3 this constraint is relaxed and the χ_{ik0} through χ_{ikv} parameters are estimated. The starting estimates for the legislators are constrained to lie within the unit hypersphere.

Given the starting estimates of the legislators, we can use the cutting plane procedure discussed in detail in Chapter 3 to get starting values for the roll call parameters. In particular, the cutting plane procedure produces estimates for the normal vectors, the $\mathbf{N_j}$'s, and the midpoints of the two outcomes, the $\mathbf{Z_j}$'s. The initial values of the (half) distances between the two outcomes, the $\boldsymbol{\delta_{jt}}$'s, can be set to 0.5 and equation (4.20) is used to get the starting values for the two outcomes. These starting values are on the normal vector line for each roll call. This process produces legislators and roll call midpoints that all meet the unit hypersphere constraint.

In Step 1 we estimate the s utility function parameters holding the legislator ideal points and roll call parameters fixed. The easiest way to do this is to set β equal to a reasonable value that can be found experimentally. (In recent DW-NOMINATE scalings of various sets of contiguous U. S. Congresses this value has been around 5.0.) Because it is very rare to estimate more than three dimensions, the second- and third-dimension weights – w_2 and w_3 – are easily found through a simple grid search over the unit square.

In Step 2 the roll call parameters are estimated with all other parameters held fixed. Because everything is held fixed, the 2s parameters, Z_{jkt} and δ_{jkt} , for each roll call

affect the log-likelihood only through the votes of the legislators on that roll call. The likelihood for the j^{th} roll call is therefore independent of the other roll calls. Technically, the second derivatives of the roll call parameters with respect to each other are all zero. This makes it fairly easy to get estimates of the 2s parameters using standard gradient methods.

For example, the BHHH algorithm developed by Berndt, Hall, Hall, and Hausman (1974) is a very simple computational approach to the problem. Let $\boldsymbol{\theta}_j$ be the 2s by 1 vector of parameters for the j^{th} roll call, $\mathbf{g}_{ij}$ be the 2s by 1 vector of the first derivatives of $\mathcal{L}$ from equation (4.24) for the i^{th} legislator on the j^{th} roll call, and let $\boldsymbol{\Omega}_j$ be the 2s by 2s matrix that is the sum of the outer products of $\mathbf{g}_{ij}$; that is:

$$\boldsymbol{\Omega}_j = \sum_{i=1}^{p} \mathbf{g}_{ij}\mathbf{g}_{ij}'$$

The update (hill climbing) formula for the parameters of the j^{th} roll call is:

$$\boldsymbol{\theta}_{j(new)} = \boldsymbol{\theta}_{j(old)} + \alpha \boldsymbol{\Omega}_j^{-1}\mathbf{g}_{ij} \qquad (4.26)$$

where α is the step size for the gradient. In normal circumstances at convergence, the matrix $\boldsymbol{\Omega}_j$ will closely approximate the *negative* of the matrix of second derivatives (the Hessian matrix). Consequently, at convergence, the diagonal of the inverse of $\boldsymbol{\Omega}_j$ yields the variances of the parameters.

If the midpoint of the roll call wanders outside the unit hypersphere, it is constrained to lie on the rim. Technically, if this constraint is invoked it means that *the log-likelihood is not necessarily at a maximum*! As a practical matter, however, this is usually not much of a problem, because roll calls that end up with their midpoints on the rim of the space are usually very lopsided, with more than 90 percent of the legislators in

the majority. Consequently, the roll call will appear to "fit" well because most of the legislators are being correctly classified, and the roll call will tend to have a high log-likelihood.

In Step 3 the parameters for the legislator ideal points – the χ_{ik0} through χ_{ikv} – are estimated with the roll call and utility parameters held fixed. Here we have sv parameters to estimate using the legislator's entire voting history. In terms of the BHHH framework, let $\boldsymbol{\theta}_i$ be the sv by 1 vector of parameters for the i^{th} legislator, let $\mathbf{g}_{ijt}$ be the sv by 1 vector of the first derivatives of $\mathcal{L}$ from equation (4.24) for the i^{th} legislator on the j^{th} roll call at time t, and let $\boldsymbol{\Omega}_i$ be the sv by sv matrix that is the sum of the outer products of $\mathbf{g}_{ijt}$; that is:

$$\boldsymbol{\Omega}_i = \sum_{t=1}^{T_i} \sum_{j=1}^{q_t} \mathbf{g}_{ijt} \mathbf{g}'_{ijt}$$

The update (hill climbing) formula for the parameters of the i^{th} legislator is:

$$\boldsymbol{\theta}_{i(new)} = \boldsymbol{\theta}_{i(old)} + \alpha \boldsymbol{\Omega}_i^{-1} \mathbf{g}_{ijt} \qquad (4.27)$$

If the constant point for the legislator -- $\boldsymbol{\chi}_{i0}$ -- wanders outside the unit hypersphere, it is constrained to lie on the rim. Again, as with the roll call midpoint constraint, *the log-likelihood is not necessarily at a maximum*! As a practical matter, however, this is usually not much of a problem either, because legislators near the rim of the space are the most ideologically rigid members and they tend to fit the model very well. The log-likelihood is usually very high for these legislators.

The three steps outlined above are repeated until the log-likelihood stops improving. At that point all the sets of parameters are reproducing each other. Within the roll call and legislator steps, each roll call or legislator is estimated while all the

remaining parameters are held fixed. The maximum of the likelihood function is found for the parameters being estimated, conditioned on all the remaining parameters being held fixed. The algorithm always moves uphill in a city-block-like fashion. At convergence, every parameter is at a maximum, conditioned on all the remaining parameters being held fixed and conditioned on the unit hypersphere constraint. Convergence to the true *global* maximum cannot be guaranteed, but the *conditional global maximum* that the algorithm reaches is almost certainly close to the overall global maximum.

The Quadratic-Normal (QN) Model

The multidimensional quadratic utility model (QN) is based on the geometry shown in Chapter 2 and uses the Optimal Classification (OC) method discussed in depth in Chapter 3 as its foundation. I developed QN in 1999 when I realized that the geometry of the quadratic utility model was identical to the geometry in OC. All that was required to make the leap to a probabilistic model was the signed distance parameter γ_j and, optionally, the legislator specific variance σ_i^2.

QN is not a dynamic program. It is designed to analyze a single legislative session, although, in principle, legislator ideal points could be estimated using the DW-NOMINATE framework outlined above.

To recap equation (4.12), the probability of the observed choice of the i^{th} legislator on the j^{th} roll call is:

$$\Phi\left[\frac{2\gamma_j(c_j - w_i)}{\sigma_i}\right]$$

and the log-likelihood is given by equation (4.16):

$$\mathcal{L} = \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{\tau=1}^{2} C_{ij\tau} \ln P_{ij\tau}$$

The total number of parameters to be estimated is:

$$q(s + 2) + p(s + 1)$$

For each roll call, $\mathbf{N_j}$, γ_j, and c_j are estimated for a total of $q(s + 2)$ parameters and for each legislator $\mathbf{X_i}$ and σ_i are estimated for a total of $p(s + 1)$. As I explained above, technically because $\mathbf{N_j}$ is a unit-length vector it is identified by the s-1 angles. But it is much easier to work directly with the normal vector than with the angles, and furthermore, the unit length constraint is in reality a "parameter" in terms of its implementation in computer code! Consequently, I will assume s+2 rather than s+1 parameters per roll call.

QN also uses the unit hypersphere constraint. Legislator ideal points – the $\mathbf{X_i}$'s -- and the projected roll call midpoints – the c_j's – are constrained to lie within the unit hypersphere. In addition, in the event of a perfect roll call, the signed distance is not allowed to "explode". I impose the constraint that the signed distance must be less than 10 units in absolute value, $|\gamma_j| < 10$. This has no real effect on the estimation, because the corresponding probabilities of the legislators are essentially 1.0. A similar constraint is imposed on the σ_i's, but it is very rarely invoked, because "perfect" voting by legislators almost never occurs, and when it does it is because a legislator votes only 20 times or so.

The parameters of the QN model are estimated in the following four-step process:

Step 0: Generate Starting Estimates for the Legislator Ideal Points, X_{ik} , the normal vectors, N_{jk} , and the projected roll call midpoints, c_j

Step 1a: Estimate the signed distance parameters, the γ_j , with the legislator ideal points, variances, and roll call normal vectors and projected midpoints held fixed

Step 1b: Estimate the Legislator Variance parameters, the σ_i , with the legislator ideal points and roll call parameters held fixed

Go to Step 1a Until Convergence

Step 2: Estimate the Legislator Ideal Points, the X_{ik} , with the legislator variances and roll call parameters held fixed

Step 3: Estimate the Projected Roll Call Midpoints, the c_j , with the normal vectors, signed distances, and legislator parameters held fixed

Step 4: Estimate the Normal Vectors, the N_{jk} , with the projected midpoints, signed distances, and legislator parameters held fixed

Go to Step 1a

I will discuss each step in turn.

Step 0 is very similar to that of DW-NOMINATE. Initial values for the legislator ideal points are produced from an eigenvalue-eigenvector decomposition of the double-centered matrix of transformed agreement scores. This produces the p by s matrix of legislator starting coordinates, $\tilde{\mathbf{X}}$, shown in equation (4.25). This initial matrix is then used as the starting coordinates for OC. OC is then run to convergence to produce the starting values of the legislator coordinates, the normal vectors, and the projected roll call midpoints. This process produces excellent starting values for the parameters because the geometry of QN is identical to that of OC.

In Step 1 the signed distances and the legislator variances are estimated in two sub-steps. This estimation is possible because, as I noted above, for each choice, γ_j and σ_i enter as the *ratio* $\frac{2\gamma_j}{\sigma_i}$. Given this fact, a simple way to start the process is to set all the σ_i equal to 1 and estimate the γ_j's with a simple grid search. With all the other parameters held fixed, the maximum log-likelihood of each of the γ_j's is easily found. Given these γ_j's, the σ_i's are also estimated with a simple grid search. With all the other parameters held fixed, the maximum log-likelihood of each of the σ_i's is also easily found. This process continues until there is no meaningful improvement in the log-likelihood. In practice, this takes no more than three repetitions.

In step 2 new legislator ideal points, the $\mathbf{X_i}$'s, are estimated within the BHHH framework discussed above. Specifically, let $\boldsymbol{\theta_i}$ be the s by 1 vector of parameters for the i^{th} legislator, $\mathbf{g_{ij}}$ be the s by 1 vector of the first derivatives of $\mathcal{L}$ from equation (4.16) for the i^{th} legislator on the j^{th} roll call, and let $\boldsymbol{\Omega_i}$ be the s by s matrix that is the sum of the outer products of $\mathbf{g_{ij}}$; that is:

$$\boldsymbol{\Omega_i} = \sum_{j=1}^{q} \mathbf{g_{ij}} \mathbf{g_{ij}}'$$

The update (hill climbing) formula for the parameters of the i^{th} legislator is:

$$\theta_{i(new)} = \theta_{i(old)} + \alpha \boldsymbol{\Omega_i^{-1}} \mathbf{g_{ij}}$$

If the legislator's ideal point wanders outside the unit hypersphere, it is constrained to lie on the rim. Again, as with DW-NOMINATE, this means that the log-likelihood is not necessarily at a maximum, but this is not a serious problem because this

type of legislator is almost always very ideologically rigid and has a very high log-likelihood.

In step 3 the projected roll call midpoints, the c_j , are estimated with all other parameters held fixed. The c_j's are easily estimated through a simple grid search because they must lie on the normal vector line between -1 and +1.

In step 4 we estimate new normal vectors, the $\mathbf{N_j}$, using standard gradient techniques with the constraints that $\mathbf{N_j'N_j} = 1$ and that $\mathbf{Z_j} = c_j\mathbf{N_j}$. In other words, the point defined by the end of the normal vector is moved along the surface of the unit hypersphere with the position of the projected midpoint held fixed on the normal vector as it is moved. Geometrically, this is equivalent to moving the cutting plane rigidly through the space as its normal vector is moved. All other parameters are held fixed. A variant of the BHHH method outlined above is used to find the new normal vectors. The unit length constraint for the normal vector is imposed at every step so that the log-likelihood is computed on the surface of the unit hypersphere.

In one dimension, given a joint rank ordering of the legislators and roll call midpoints from the classification algorithm, step 4 is not necessary and the legislator coordinates in step 2 can be found through a simple grid search. In practice only 3 overall passes through steps 1 to 4 are required for the QN algorithm to converge.

An attractive feature of the QN model is that *if the ideal points were known*, the roll call parameters could be estimated by probit regression rather than steps 3 and 4. A probit specification is logically consistent with the assumptions of the ideal point estimator. For example, in a probit regression of a particular roll call vote on party and the estimated ideal point, the null model in which only the ideal point mattered would be

correctly specified. As I discussed in Chapter 2, however, given that probit becomes unreliable at very low levels of error, some safeguards have to be built in to catch perfect and near perfect roll calls.[13]

The four steps above are repeated until the log-likelihood stops improving. At that point all sets of parameters are reproducing each other. Within each step the maximum of the likelihood function is found for the parameter being estimated conditioned on all the remaining parameters being held fixed. The algorithm always moves uphill in a city-block-like fashion. At convergence, every parameter is at a maximum conditioned on all the remaining parameters being held fixed and conditioned on the unit hypersphere constraint.[14]

Statistical Issues

Two important statistical issues affect the parameters estimated by DW-NOMINATE and QN: standard errors and the "incidental parameters" problem that may produce bias in the parameter estimates.

In classical maximum likelihood, the maximum of the log likelihood, $\mathcal{L}$, is found by taking the first derivatives with respect to all the parameters, setting them equal to zero, and then solving the set of equations for the parameter values. These become the maximum likelihood estimates. For the solution to be a maximum, the matrix of second derivatives must be negative definite. For example, suppose there are 100 legislators and 500 roll calls. QN would require the estimation of 2,300 parameters, so that the vector of parameters, $\boldsymbol{\theta}$, would be a 2,300 by 1 vector. In the BHHH framework, the vector of first

derivatives, $\mathbf{g_{ij}}$, for the i^{th} legislator on the j^{th} roll call would be also 2,300 by 1, and the matrix of the sum of the outer products would be:

$$\mathbf{\Omega} = \sum_{i=1}^{p}\sum_{j=1}^{q}\mathbf{g_{ij}}\mathbf{g_{ij}}'$$

so that $\mathbf{\Omega}$ would be a 2,300 by 2,300 matrix. As I noted above, in normal circumstances at convergence, the matrix $\mathbf{\Omega}$ will closely approximate the *negative* of the matrix of second derivatives (the Hessian matrix). Consequently, at convergence, $\mathbf{\Omega}$ will be a positive definite matrix and the variances of the estimated parameters are on the diagonal of the inverse of $\mathbf{\Omega}$.

Unfortunately, the presence of constraints and the sheer size of $\mathbf{\Omega}$ make the classical approach of obtaining standard errors infeasible.

Even if the constraints were not present, the maximum likelihood approaches outlined above have no way of getting around the "incidental parameters" problem. Simply stated, parameters are estimated for the rows of the data (usually the legislators are the rows) and the columns of the data (usually the roll calls). Every time a roll call is added to the data matrix, 2s additional parameters in DW-NOMINATE and s+2 additional parameters in QN must be estimated. If a legislator is added, then s additional parameters in DW-NOMINATE and s+1 (if a legislator-specific variance is estimated) additional parameters in QN must be estimated. That is, as the data grow so does the number of parameters. Consequently, the standard proof of the consistency of maximum likelihood does not apply (Neyman and Scott, 1948).

Haberman (1977) analyzed the simple version of the item response model developed by Rasch (1961) and was able to show consistency with some restrictive conditions. The simple item response model is mathematically equivalent to the basic

spatial model if legislators have quadratic utility functions with additive random error (Ladha, 1991; Londregan, 2000; Clinton, Jackman, and Rivers, 2004). Haberman (1977) shows that as both the number of legislators and the number of roll calls go to infinity with the restriction that log(q)/p go to infinity as well, then the simple Rasch model is consistent.

The most important recent work on consistency is by Londregan (2000). He links the psychometrics testing literature with the spatial theory of legislative voting and derives important statistical results about the parameters of the spatial model. In particular, when the preferential choices are *nominal*, Londregan shows that consistency in its usual statistical sense does not hold. With nominal choices standard maximum likelihood estimators that attempt *simultaneously* to recover legislators' ideal points and roll call parameters inherit the "granularity" of the choice data and so cannot recapture the underlying *continuous* parameter space.[15] If legislators could report continuous "feeling thermometer" scores instead of just Yea/Nay, this source of inconsistency would disappear.

Unfortunately, in the real world of data we are faced with finite sample sizes so that there is no simple way of ensuring that our estimates of legislator ideal points and roll call parameters are unbiased and no simple way of solving for standard errors. Consequently, there is no substitute for Monte Carlo analysis in establishing the quality of the estimated parameters.

Howard Rosenthal and I performed a large number of simulation studies of our NOMINATE model and found that it performed very well (Poole and Rosenthal, 1991; 1997). Our approach was to create artificial data that met all the assumptions of the

model and then test how accurate the method was in reproducing the true legislator ideal points and roll call parameters. We found that when the number of legislators was 50 or more with 100 or more roll calls, the recovery of the true parameters by NOMINATE was very good. We concluded that bias was not a serious problem with reasonably sized roll call matrices.

Obtaining standard errors was a harder problem. We decided to use Efron's (1979) nonparametric bootstrap procedure to obtain standard errors for the legislator ideal points. We sampled roll calls with replacement holding the number of legislators fixed and created 50 matrices.[16] We then ran NOMINATE on each of the 50 roll call matrices and computed standard errors for the legislators from the 50 sets of ideal points.

This technique produced nice looking standard errors, but we were never completely satisfied with it because it could be used to get the standard errors of only *one set of parameters*. You had to sample either the rows of the matrix or the columns of the matrix.

In the past few years two approaches have emerged that appear to have successfully solved the standard error problem. The first is based on *Markov Chain Monte Carlo* (MCMC) simulation (Metropolis and Ulam, 1949; Hastings, 1970; Geman and Geman, 1984; Gelfand and Smith, 1990; Gelman, 1992) within a Bayesian framework (Gelman, Carlin, Stern, and Rubin, 2000; Gill, 2002) and was introduced into Political Science by Simon Jackman (2000a; 2000b; 2001; Clinton, Jackman, and Rivers, 2004) and Andrew Martin and Kevin Quinn (Schofield, Martin, Quinn, and Whitford, 1998; Quinn, Martin, and Whitford, 1999; Martin and Quinn, 2002; Quinn and Martin, 2002; Martin, 2003; Quinn, 2004). I will discuss this innovative approach in the next

section. The second is based on the parametric bootstrap (Efron, 1979; Efron and Tibshirani, 1993). Jeff Lewis conceived this approach, and it is discussed in depth by Lewis and Poole (2004). I will outline it after I discuss Bayesian simulation.

The Bayesian Simulation Approach

The basic idea behind Bayesian simulation is to "illuminate" a probability density function of the parameters – the posterior distribution – through a "random tour" of the parameter space. At each point of the random tour the value of the parameter value is recorded so that when the tour is finished – typically hundreds of thousands or millions of points in the parameter space are visited – the shape of the distribution over the parameter space is known with some accuracy. At the end of the tour the means and standard errors of the parameters can be directly calculated from the tour record.

The random tour is generated using MCMC methods. As Shawn Trier and Simon Jackman (2003, p.10) note, "the virtue of MCMC methods is that (subject to regularity conditions) they produce random tours that visit locations in the parameter space with frequency proportional to their posterior probability." With respect to estimation of spatial voting models, the advantage of this approach is that it produces estimates of the legislator ideal points and roll call parameters and all of the standard errors as well. In addition, it is also possible to estimate and do inference about other quantities of interest such as the median or any function of the model parameters (Jackman, 2000b). The disadvantage is that the MCMC approach can consume a very large amount of computer time because the bigger the roll call matrix, the larger the number of parameters, and the longer it takes to do a comprehensive random tour of the parameter space.

Technically, the Bayesian simulation approach works with a *posterior* probability distribution. The intuition behind this can be seen by looking at the simple formulas for conditional probability. Let θ and Y be two events, then in classical probability theory:

$$P(\theta \mid Y) = \frac{P(\theta \cap Y)}{P(Y)} \quad \text{and} \quad P(Y \mid \theta) = \frac{P(\theta \cap Y)}{P(\theta)}$$

hence

$$P(\theta \cap Y) = P(\theta \mid Y)P(Y) = P(Y \mid \theta)P(\theta)$$

And

$$P(\theta \mid Y) = \frac{P(Y \mid \theta)P(\theta)}{P(Y)}$$

In the Bayesian framework, Y is the observed data, the θ are the parameters, P(Y | θ) is the joint distribution of the sample, P(θ) is the *prior distribution* of the parameters, P(Y) is the marginal distribution of the sample, and P(θ | Y) is the *posterior distribution*. Because P(Y) is a constant, the posterior distribution is proportional to the product of the joint distribution of the sample (which is proportional to the likelihood function) and the prior distribution; that is:

$$P(\theta \mid Y) \propto P(Y \mid \theta)P(\theta) \qquad (4.28)$$

In the spirit of Bayes, the researcher specifies the prior distribution of the parameters; that is, P(θ). Typically the prior distribution is chosen so that when it and the joint distribution of the sample are multiplied, the general form of the posterior distribution is known. For example, if the joint distribution of the sample is normally distributed and the prior distribution is assumed to be a normal distribution, the posterior distribution will be a normal distribution.

If the parameters can be partitioned into k subsets, that is, $\boldsymbol{\theta_1}$, $\boldsymbol{\theta_2}$, …, $\boldsymbol{\theta_k}$, then the MCMC random tour through the parameter space is made much more tractable because it can be "generated by successively sampling from the conditional distributions that together characterize the joint posterior density" (Trier and Jackman, 2003, p.10). Technically, this is *alternating conditional sampling* or the *Gibbs sampler* (Geman and Geman, 1984; Gelfand and Smith, 1990).[17] For example, at the h^{th} iteration the MCMC sampling procedure is:

1. Sample $\theta_{1i}^{(h)}$ from $\mathbf{g}_{\boldsymbol{\theta}_1}\left(\boldsymbol{\theta}_{1i} \mid \boldsymbol{\theta}_2^{(h-1)}, \boldsymbol{\theta}_3^{(h-1)}, ..., \boldsymbol{\theta}_k^{(h-1)}, \mathbf{Y}\right)$, i=1,…,$n_1$

2. Sample $\theta_{2i}^{(h)}$ from $\mathbf{g}_{\boldsymbol{\theta}_2}\left(\boldsymbol{\theta}_{2i} \mid \boldsymbol{\theta}_1^{(h-1)}, \boldsymbol{\theta}_3^{(h-1)}, ..., \boldsymbol{\theta}_k^{(h-1)}, \mathbf{Y}\right)$, i=1,…,$n_2$

……

k. Sample $\theta_{ki}^{(h)}$ from $\mathbf{g}_{\boldsymbol{\theta}_k}\left(\boldsymbol{\theta}_{ki} \mid \boldsymbol{\theta}_1^{(h-1)}, \boldsymbol{\theta}_2^{(h-1)}, ..., \boldsymbol{\theta}_{k-1}^{(h-1)}, \mathbf{Y}\right)$, i=1,…,$n_k$ (4.29)

where $\sum_{j=1}^{k} n_j$ is the total number of parameters.

When this framework is applied to roll call data, $\mathbf{Y}$ is the p by q matrix of choices and $\boldsymbol{\theta}$ is the vector of legislator ideal points and roll call parameters. With quadratic utility, normally distributed error, and normally distributed prior distributions, the conditional distributions are also normally distributed. This Bayesian QN framework is equivalent to the standard two parameter item response model widely used in educational testing (see Chapter 2).

Specifically, using equations (4.5) to (4.7) let

$$\alpha_j = \left(\sum_{k=1}^{s} O_{jkn}^2 - \sum_{k=1}^{s} O_{jky}^2\right) = 2\gamma_j \mathbf{Z_j}'\mathbf{N_j} \quad \text{and}$$

$$\boldsymbol{\beta}_j = \begin{bmatrix} -2(O_{j1n} - O_{j1y}) \\ -2(O_{j2n} - O_{j2y}) \\ \vdots \\ -2(O_{jsn} - O_{jsy}) \end{bmatrix} = -2\gamma_j \mathbf{N_j} \qquad (4.30)$$

This transformation allows the difference between the latent utilities for Yea and Nay to be written in the same form as the item response model; namely:

$$y^*_{ij} = U_{ijy} - U_{ijn} = \alpha_j + \mathbf{X_i'\beta_j} + \varepsilon_{ij} \qquad (4.31)$$

where y^*_{ij} is the difference between the latent utilities and

$$\varepsilon_{ij} = \varepsilon_{ijn} - \varepsilon_{ijy} \sim N(0, 1)$$

If the latent utility differences were observed, then the joint distribution of the sample would be:

$$\mathbf{f(Y^* \mid \alpha, \beta, X)} = \prod_{i=1}^{p}\prod_{j=1}^{q} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y^*_{ij} - \alpha_j - X_i'\beta_j)^2} \qquad (4.32)$$

Where $\mathbf{Y^*}$ is the p by q matrix of latent utility differences, $\boldsymbol{\alpha}$ is a q by 1 vector of the $2\gamma_j\mathbf{Z_j'N_j}$ terms, $\boldsymbol{\beta}$ is q by s matrix of the $-2\gamma_j\mathbf{N_j}$ terms, and $\mathbf{X}$ is the p by s matrix of legislator ideal points.

Unfortunately, the latent utility differences are not observed and we do not have any simple expression for $\mathbf{f(Y \mid \alpha, \beta, X)}$. However, $\mathbf{f(Y \mid \alpha, \beta, X)}$ is proportional to the likelihood function, $\mathbf{L(\alpha, \beta, X \mid Y)}$ (equation [4.15]), and this fact allows the construction of the conditional distributions for the posterior distribution. Specifically, let:

$$y_{ij} = \begin{cases} 1 \text{ (Yea) if } y^*_{ij} > 0 \\ 0 \text{ (Nay) if } y^*_{ij} \le 0 \end{cases} \text{ so that } \begin{cases} \mathbf{P(y^*_{ij} > 0) = \Phi(\alpha_j + X_i'\beta_j)} \\ \mathbf{P(y^*_{ij} \le 0) = 1 - \Phi(\alpha_j + X_i'\beta_j)} \end{cases}$$

If the y_{ij} are independent Bernoulli random variables, that is:

$$\mathbf{f(y_{ij} \mid \alpha_j,\ \beta_j,\ X_i) \sim Bernoulli(\Phi(\alpha_j + X_i'\beta_j))}$$

then

$$\mathbf{f(Y \mid \alpha, \beta, X)} \propto \prod_{i=1}^{p}\prod_{j=1}^{q} f(y_{ij} \mid \alpha_j,\ \beta_j,\ X_i) = \mathbf{L(\alpha, \beta, X \mid Y)} =$$

$$\prod_{i=1}^{p}\prod_{j=1}^{q} [\Phi(\alpha_j + X_i'\beta_j)]^{y_{ij}} [1 - \Phi(\alpha_j + X_i'\beta_j)]^{(1-y_{ij})} \quad (4.33)$$

Note that the reason that the joint distribution of the sample is only *proportional* to the likelihood function is that $\mathbf{L(\alpha, \beta, X \mid Y)}$ in this case is *not a proper probability distribution*. That is, the multidimensional integral of $\mathbf{L(\alpha, \beta, X \mid Y)}$ is not equal to 1. If the value of that integral were known, then dividing the likelihood function by it would yield the joint distribution of the sample.

Let the prior distribution for the legislator ideal points be:

$$\mathbf{X_i} \sim N(\mathbf{0}, \mathbf{I_s}) = \xi(\mathbf{X_i}) \quad (4.34)$$

Where $\mathbf{0}$ is a s length vector of zeroes and $\mathbf{I_s}$ is an s by s identity matrix. The prior distribution for the roll call outcome parameters is:

$$\begin{bmatrix} \alpha_j \\ \beta_j \end{bmatrix} \sim N(\mathbf{b_0}, \mathbf{B_0}) = \xi(\alpha_j, \boldsymbol{\beta_j}) \quad (4.35)$$

where $\mathbf{b_0}$ is an s+1 length vector and $\mathbf{B_0}$ is a s+1 by s+1 variance-covariance matrix. Clinton, Jackman, and Rivers (2004) set $\mathbf{b_0}$ to a vector of zeroes and $\mathbf{B_0}$ to ηI_{s+1} where η is a large positive constant (typically 25).

With these assumptions, the posterior distribution for the Bayesian QN model is proportional to the product of the joint distribution of the sample and the prior distributions; namely:

$$\xi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{X} \mid \mathbf{Y}) \propto \mathbf{f(Y \mid \alpha, \beta, X)}\xi(\mathbf{X})\xi(\boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \mathbf{L(\alpha, \beta, X \mid Y)}\xi(\mathbf{X})\xi(\alpha, \boldsymbol{\beta}) \quad (4.36)$$

Because the latent utility differences are used in the sampling procedure, it is useful to rewrite the posterior distribution to reflect this; namely

$$\xi(\alpha, \beta, X \mid Y) = \int \xi(Y^*, \alpha, \beta, X \mid Y) dY^*$$

Martin and Quinn (2001, p. 12) note that "conditioning on the latent utility differences allows us to recast the item response model for a dichotomous response variable as a factor analysis model for a (latent) continuous response. The integration over $[Y^*]$ is easily accomplished in the simulation…" In other words, we can pretend that we have all the y^*_{ij}'s from equation (4.32) so that we do not have to use equation (4.33). The sampling is from the conditional distributions of the joint posterior distribution $\xi(Y^*, \alpha, \beta, X \mid Y)$ and the draws of the y^*_{ij} can be ignored (Albert and Chibb, 1993). What this means is that the Markov chain of the roll call and legislator parameters (under certain regularity conditions) converges to the joint posterior distribution that we are interested in, namely, equation (4.36).[18]

The conditional distributions for $\xi(Y^*, \alpha, \beta, X \mid Y)$ that implement equation (4.36) are:

$$1)\quad g_{y^*_{ij}}\left(y^*_{ij} \mid \alpha_j, \beta_j, X_i, y_{ij}\right) = \begin{cases} N_{[0,\infty)}(\alpha_j + X_i'\beta_j, 1) \text{ if } y_{ij} = \text{Yea} \\ N_{(-\infty,0]}(\alpha_j + X_i'\beta_j, 1) \text{ if } y_{ij} = \text{Nay} \\ N_{(-\infty,\infty)}(\alpha_j + X_i'\beta_j, 1) \text{ if } y_{ij} = \text{Missing} \end{cases}$$

where the subscript on the normal distribution indicates the range. For Yea and Nay the normal is truncated as indicated and the missing data is sampled over the entire real line.

$$2)\quad g_{\alpha,\beta}\left(\alpha_j, \beta_j \mid Y^*_j, X, Y_j\right) = N(\nu_j, \Xi_j)$$

where Y_j^* and $\mathbf{Y_j}$ are the j^{th} columns of $\mathbf{Y}^*$ and $\mathbf{Y}$, respectively, $\boldsymbol{\nu_j}$ is an s+1 length vector of means, and $\boldsymbol{\Xi_j}$ is the s+1 by s+1 variance-covariance matrix. Specifically,

$$\boldsymbol{\nu_j} = \boldsymbol{\Xi_j}\left[\mathbf{X}^{*\prime}\mathbf{Y}_j^* + \mathbf{B}_0^{-1}\mathbf{b}_0\right]$$

and

$$\boldsymbol{\Xi_j} = \left[\mathbf{X}^{*\prime}\mathbf{X}^* + \mathbf{B}_0^{-1}\right]^{-1}$$ where $\mathbf{X}^*$ is the p by s+1 matrix of legislator ideal points bordered by ones; that is, $\mathbf{X}^* = \mathbf{[J_p|X]}$ and $\mathbf{J_p}$ is a p length vector of ones.

3) $$\mathbf{g_X}\left(\mathbf{X_i} \mid \alpha, \beta, \mathbf{Y_i^*}, \mathbf{Y_i}\right) = \mathbf{N(t_i, T_i)}$$

where Y_i^* and $\mathbf{Y_i}$ are the i^{th} rows of $\mathbf{Y}^*$ and $\mathbf{Y}$, respectively, $\mathbf{t_i}$ is an s length vector of means, and $\mathbf{T_i}$ is the s by s variance-covariance matrix. Specifically,

$$\mathbf{t_i} = \mathbf{T_i}\left[\beta'(\alpha - \mathbf{Y_i^*})\right]$$

and

$$\mathbf{T_i} = \left[\beta'\beta + \mathbf{I_s}\right]^{-1}$$

Note that if a more general prior distribution is used for the legislator ideal points, namely:

$$\mathbf{X_i} \sim \mathrm{N}(\mathbf{t_0}, \mathbf{T_0})$$

Where $\mathbf{t_0}$ is a s length vector of means and $\mathbf{T_0}$ is a s by s variance-covariance matrix, then the expressions for $\mathbf{t_i}$ and $\mathbf{T_i}$ are:

$$\mathbf{t_i} = \mathbf{T_i}\left[\beta'(\alpha - \mathbf{Y_i^*}) + \mathbf{T_0^{-1}t_0}\right]$$

and

$$\mathbf{T_i} = \left[\beta'\beta + \mathbf{T_0}\right]^{-1}$$

At the h^{th} iteration the MCMC sampling procedure outlined in equation set (4.29) for these conditional distributions is:

1) Sample $y_{ij}^{*(h)}$ from $\mathbf{g}_{y_{ij}^*}\left(\mathbf{y}_{ij}^* \mid \boldsymbol{\alpha}_j^{(h-1)}, \boldsymbol{\beta}_j^{(h-1)}, \mathbf{X}_i^{(h-1)}, \mathbf{y}_{ij}\right)$, i=1,…,p , j=1,…,q

2) Sample $\boldsymbol{\alpha}_j^{(h)}, \boldsymbol{\beta}_j^{(h)}$ from $\mathbf{g}_{\alpha,\beta}\left(\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j \mid \mathbf{Y}_j^{*(h-1)}, \mathbf{X}^{(h-1)}, \mathbf{Y}_j\right)$, j=1,…,q (4.37)

3) Sample $\mathbf{X}_i^{(h)}$ from $\mathbf{g}_X\left(\mathbf{X}_i \mid \boldsymbol{\alpha}^{(h-1)}, \boldsymbol{\beta}^{(h-1)}, \mathbf{Y}_i^{*(h-1)}, \mathbf{Y}_i\right)$, i=1,…,p

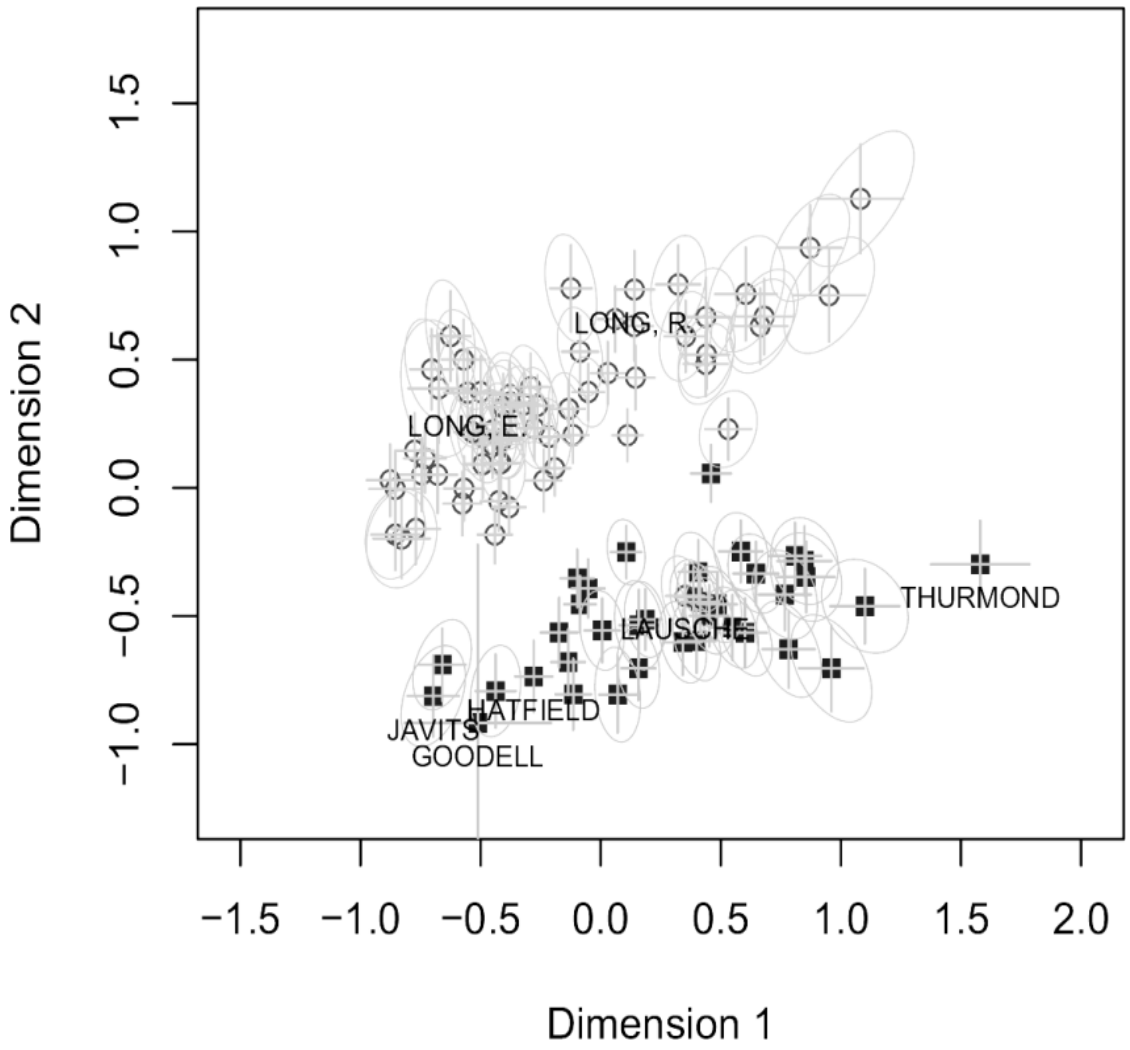
Sampling from these conditional distributions is straightforward because they are all normal distributions with known means and variances. If the conditional distributions are non-standard then a sampling method such as the Metropolis-Hastings algorithm must be used (Metropolis and Ulam, 1949; Hastings, 1970; Gelman, Carlin, Stern, and Rubin, 2000, ch. 11; Gill, 2002, ch. 9).

Let $\boldsymbol{\zeta}^{(h)}$ denote all the parameters sampled during the h^{th} iteration through equation set (4.37), that is; $\boldsymbol{\zeta}^{(h)} = \left\{\boldsymbol{\alpha}^{(h)}, \boldsymbol{\beta}^{(h)}, \mathbf{X}^{(h)}\right\}$. The sequence $\boldsymbol{\zeta}^{(1)}, \boldsymbol{\zeta}^{(2)}, \boldsymbol{\zeta}^{(3)}, \ldots$ is a Markov Chain that converges to the posterior distribution, $\boldsymbol{\xi}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{X} \mid \mathbf{Y})$ (equation [4.36]), given certain regularity conditions.[19] (Note that $\mathbf{Y}^*$ is not part of the chain.) Typically, tens of thousands of samples are drawn and inference – means, medians, and confidence intervals -- are based on the final 100,000 or so samples.

This model (with suitable elaborations) has been applied to congressional voting by Clinton, Jackman, and Rivers (2004), to voting in the United Nations by Erik Voeten (2004), to voting in the Continental Congress by Joshua Clinton (2003), to voting in the Supreme Court by Andrew Martin and Kevin Quinn (2002), and to a comparison of voting in the Senate and the Supreme Court by Michael Bailey (2002), to cite just a few papers that have been generated by this approach in the past few years.[20]

Figure 4.4 shows an example of Bayesian QN applied to the 90th U.S. Senate.[21] Squares represent Democrats and circles represent Republicans. The vertical and horizontal lines through each Senator's ideal point show the 95 percent "confidence" intervals for each coordinate of a given Senator's position.[22] For most Senators, the posterior correlation between the estimated first and second dimension estimates is very low. Normal theory 95 percent "confidence" ellipses are shown for Senators whose first dimension coordinate is correlated with their second dimension coordinate greater than 0.3 in absolute value.

Figure 4.4: 90th U.S. Senate from Bayesian QN



On average the ideal points are precisely estimated. The average standard error on the first dimension is 0.05 and 0.07 on the second dimension (Lewis and Poole, 2004, p. 120). The senator with the largest standard errors around his ideal point is Charles Goodell (R-NY). He was appointed to the Senate in September 1968 to replace Robert F. Kennedy after his assassination that June. Goodell only voted on 31 scalable roll calls (2.5 percent or more in the minority) so that his standard errors are huge.

The advantage of the Bayesian simulation approach is that it produces parameter estimates and their standard errors in one process. The disadvantage is that it is very computationally intensive. So far, all the applications have used the quadratic utility framework because it is straightforward to write out the conditional distributions in the equation system (4.37). It has not yet been applied to the normal distribution utility model. This is a difficult problem because the conditional distributions do not have a clear-cut form and would themselves have to be simulated.

The simulation approach can only improve over time. Its structure is ideally suited for massively parallel processing and computers are only going to get faster and cheaper in the coming decades. Howard Rosenthal and I had to use a Cyber 205 vector supercomputer in the 1980s to run D-NOMINATE. Its much more flexible successor, DW-NOMINATE, outlined earlier in this Chapter, will run on a personal computer.

The Parametric Bootstrap[23]

In the Bayesian MCMC framework the parameters of interest (the ideal point and roll call parameters) are treated as random variables. Subjective probability distributions are placed over the analyst's uncertainty about the unknown parameters. By Bayes' Rule, when the likelihood function is multiplied by these prior distributions the result is proportional to the posterior distribution of the parameters. The MCMC process approximates draws from the posterior distribution of the parameters. In contrast, the parametric bootstrap takes the classical statistical approach. The parameters are treated as fixed constants to be estimated and consequently, they are not given probability distributions. Rather, the underlying probability model, with its parameters set equal to

the maximum likelihood estimates, is used to generate repeated simulated samples. For each of these repeated samples, the model is refit, and the empirical distribution of these estimates across the pseudo-samples is used to estimate the relevant sampling distributions.

The parametric bootstrap is conceptually simple. In a maximum likelihood framework, the first step is to compute the likelihood function of the sample. The second step is to draw, for example, 1000 samples from the likelihood density and compute for each sample the maximum likelihood estimates of the parameters of interest. The sample variances computed from these 1000 values are the estimators of the variances of the parameters (Efron and Tibshirani, 1993, ch. 6). The parametric bootstrap assumes that the model has been correctly specified so that its assumptions are stronger than the nonparametric bootstrap.

When applied to a scaling method such as NOMINATE or QN, the first step is to run the program to convergence and then calculate the probabilities for the observed choices. This produces a legislator by roll call matrix containing the estimated probabilities for the corresponding actual roll call choices of the legislators. Note that the product of these probabilities is the likelihood; that is:

$$\mathbf{L}(\hat{\theta} \mid \mathbf{Y}) = \prod_{i=1}^{p} \prod_{j=1}^{q} \prod_{\tau=1}^{2} \hat{\mathbf{P}}_{ij\tau}^{C_{ij\tau}}$$

Where $\hat{\theta}$ is the set of all the estimated ideal point, roll call, and utility function parameters, τ is the index for Yea and Nay, $\hat{P}_{ijc}$ is the estimated probability of voting for choice τ as given by equations (4.10) or (4.13), and $C_{ij\tau} = 1$ if the legislator's actual choice is τ and zero otherwise.

To draw a random sample treat each estimated probability as a weighted coin and "flip" the coin. Draw a number from a uniform distribution over zero to one -- U(0,1) – and if the estimated probability is greater than or equal to the random draw, the sampled value is the observed choice. If the random draw is greater than the estimated probability, then the sampled value is the opposite of the observed choice; that is, if the observed choice is Yea then the sampled value is Nay. Note that this is equivalent to drawing a random sample of size one from pq separate Bernoulli distributions with parameters $\hat{P}_{ijc}$. In the context of a Bernoulli distribution the two outcomes are either "success" or "failure". Here a "success" is the observed choice and "failure" is the opposite choice.

The sample roll call matrix is then analyzed by NOMINATE or QN. Repeat this process 1000 times and calculate the variances of the legislator ideal points using the 1000 estimated bootstrap configurations.

Technically, let π be a random draw from U(0,1). The sample rule is:

If the observed choice is Yea (Nay):

$$\text{If } \pi \leq \hat{P}_{ijc} \text{ then the sample value, } \hat{C}_{ij}, \text{ is Yea (Nay)} \qquad (4.39)$$

$$\text{If } \pi > \hat{P}_{ijc} \text{ then the sample value, } \hat{C}_{ij}, \text{ is Nay (Yea)}$$

This technique allows the underlying uncertainty to propagate through to all the estimated parameters. To see this, note that as $\hat{P}_{ijc} \to 1$, then $\hat{C}_{ij} \to C_{ij}$, that is, sample choices become the observed choices so that the bootstrapped variances for the parameters of the model go to zero. If the fit of the model is poor, for example, if the $\hat{P}_{ijc}$ are between .5 and .7, then the bootstrapped variances for the parameters will be large.

Let $\hat{\mathbf{X}}$ be the p by s matrix of legislator coordinates estimated by either W-NOMINATE or QN. Let h=1,…,m be the number of bootstrap trials and let $\mathbf{X_h}$ be the p by s matrix of legislator coordinates estimated on the h^{th} bootstrap trial. The legislator and roll call coordinates are only identified up to an arbitrary rotation in the s-dimensional space. This arbitrary rotation must be removed to ensure that the bootstrapping process produces accurate estimates of the standard deviations of the parameters. In particular, let:

$$\hat{\mathbf{X}} = \mathbf{X_h V} + \mathbf{E} \qquad (4.40)$$

$\mathbf{V}$ is an s by s matrix such that $\mathbf{V'V} = \mathbf{VV'} = I_s$, I_s is an s by s identity matrix, and $\mathbf{E}$ is a p by s matrix of error. In psychometrics, equation (4.40) is known as the *orthogonal procrustes problem*. Schonemann's (1966) solution is used to remove the arbitrary rotation, $\mathbf{V}$. Note that $\mathbf{X_h}$ is being *rigidly rotated* so that the estimated points are not being altered vis a vis one another in any way. Consequently, in the formulas for the mean and standard deviation below I use $\mathbf{X_h}$ to denote the h^{th} bootstrap trial to reduce notational clutter.

The mean legislator ideal point on the k^{th} dimension is:

$$\bar{X}_{ik} = \frac{\sum_{h=1}^{m} X_{hik}}{m} \qquad (4.41)$$

where X_{hik} is the estimated coordinate on the h^{th} trial. The corresponding standard deviation is:

$$s\{X_{ik}\} = \sqrt{\frac{\sum_{h=1}^{m} (X_{hik} - \hat{X}_{ik})^2}{m-1}} \qquad (4.42)$$

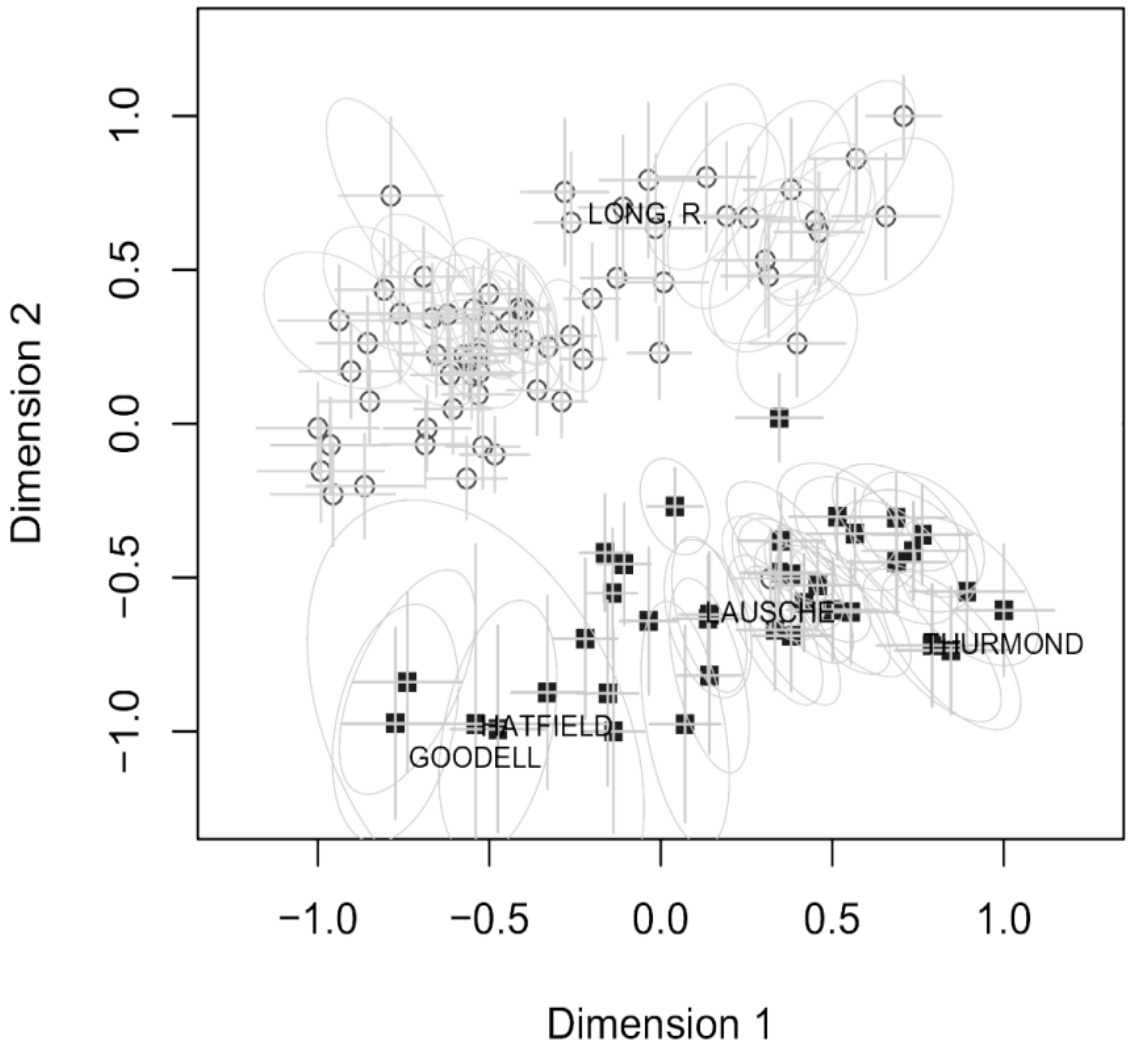
where $\hat{X}_{ik}$ is the coordinate estimated by W-NOMINATE or QN.

Jeff Lewis and I (2004) take a conservative approach and use the estimated coordinate, $\hat{X}_{ik}$, rather than the mean of the bootstrap trials, $\bar{X}_{ik}$, as our "sample mean" in our calculation of the standard deviation. This inflates the standard deviations somewhat but we feel it is better to err on the safe side and not *underreport* the standard deviations.

This technique is superior to the nonparametric bootstrap approach that Howard Rosenthal and I used. The parametric bootstrap allows the uncertainty to be reflected in *all of the choices* and therefore to pass through to the estimates of *all of the parameters*. In addition, it is simple computationally. It consists simply of adding a step at the end of the program to save the probabilities of all the choices and then looping through the scaling program m additional times.

Figure 4.5 shows an example of the parametric bootstrap of QN applied to the 90th U.S. Senate. Squares represent Democrats and circles represent Republicans. The vertical and horizontal lines through each Senator's ideal point show the 95 percent confidence intervals for each coordinate of a given Senator's position. For most Senators, the correlation between the estimated first and second dimension estimates is very low. Normal theory 95 percent confidence ellipses are shown for Senators whose first dimension coordinate is correlated with their second dimension coordinate greater than 0.3 in absolute value.

Figure 4.5: 90th U.S. Senate Bootstrapped QN



The QN ideal points are precisely estimated. The average standard error on the first dimension is 0.05 and 0.08 on the second dimension (Lewis and Poole, 2004, p. 120). The senator with the largest standard errors around his ideal point is Charles Goodell (R-NY). The Pearson correlation between the first dimension coordinates from QN and those from Bayesian QN (Figure 4.4) is 0.99 and the corresponding second dimension correlation is 0.97 (Lewis and Poole, 2004, p.117). The two methods produce essentially the same ideal points.

Conclusion

In this chapter I outlined the major probabilistic models of legislative voting. Estimation of these models is complicated, and the statistical issues are especially thorny and not fully resolved. Nevertheless, considerable progress has been made in dealing with the two major statistical issues – incidental parameters and standard errors – and it is quite likely that there will be further progress in the near future. The good news is that the major probabilistic methods (Bayesian and non-Bayesian) produce essentially the same spatial maps for large legislatures (Lewis and Poole, 2004). This is important from a scientific standpoint, because it means that the basic theoretical model of choice shared by all the estimation methods produces meaningful results.

The importance that I place on the basic theoretical model of choice (see Chapter 7) is the reason that I am agnostic when it comes to the Frequentist vs. Bayesian debate.[24] I taught statistics to MBA students for nearly 20 years, and I was invariably asked "How can I *use* this in business?" My answer was always that they should regard these methods as tools. Tools are a means to an end. You must understand how to use the tool and what its limitations are. Sometimes, indeed many times, more than one tool can do the job.

This is how I regard the methods I outlined in this chapter. If roll call voting in a legislature is highly structured, then different methods should produce very similar spatial maps. Now that two methods of obtaining good standard errors are available – MCMC simulation and the parametric bootstrap -- both should be used as checks on each other. The aim is to build a science of choice. The methods are a means to this end.

In Chapter 5 I turn to practical issues of estimation of spatial models – how to generate good starting values, how to determine dimensionality, how to interpret the

dimensions, and so on, along with some computing tricks I have used over the years that those eager readers who want to “roll their own” might find useful.

Chapter Four Notes

[1] Technically I am assuming that the error distribution is *continuous*. Consequently, I do not have to worry about the case where $\mathbf{P(}U_{ijy} - U_{ijn} = 0\mathbf{)}$.

[2] Recall from Chapter 1 that a strong regularity emerged from the work in psychology on multidimensional scaling: when people judged the similarity between stimuli or were asked to report preferences for stimuli, they appeared to be using an *exponential response function*. That is, when the actual judged similarities are graphed against some objective measure of similarity, they almost always produce an exponential-like plot (Shepard, 1987). Specifically, if the similarity measure, A, is a number between zero (most similar) and one (least similar), people tend to report e^{-kA} where k is a scaling constant. Because people are not perfect judges, when perceptual error is added to Shepard's model, the expected value of the response function is *Gaussian* – that is, e^{-kA^2} (Nosofsky, 1986, 1988; Ennis, 1988a, 1988b; Shepard, 1988a; 1988b; Ennis, Palen, and Mullen, 1988). See Chapter 7 for a more detailed discussion of the implications of these results for geometric models of choice in Political Science.

[3] The nesting idea was the product of a conversation between George Rabinowitz and Howard Rosenthal at a conference held at Carnegie-Mellon University in 1987. Merrill and Grofman (1999, chs. 1 –3) elaborate on this framework and show how the classical Downsian model and the directional model can be combined in a unified model that can

be empirically tested. Their bottom-line findings are that voter choice appears to be a *mix* of directional and proximity components and that this mix varies with the type of candidate (challenger or incumbent). The key empirical result is that voters appear to respond to candidates offering policies in their direction but reject those candidates distant from themselves.

[4] There is a large literature on (and considerable controversy about) the directional versus the proximity (traditional spatial theory of voting) model of choice in mass publics. For example, see Lewis and King (1999) and Macdonald, Rabinowitz, and Listhaug (2001) and the citations therein. The most comprehensive analysis of the two models is by Merrill and Grofman (1999).

[5] "Directional distance" as it is used in this chapter is not to be confused with the directional utility model. By directional distance I mean only that it is *signed*.

[6] Except for an added "valence" dimension, this model is identical to the one-dimensional model developed by Londregan (2000, p. 40-41). Specifically, in Londregan's notation, $g = (O_{jn} - O_{jy})$, $m = (O_{jn} + O_{jy})/2$, $x_v = x_i$, and $z = O_{jy}$ or O_{jn} depending upon which is the *proposal* (the opposite alternative is the status quo). The utility function used by Londregan is:

$$U(z, q| x_v) = (-1/2)(z - x_v)^2 + \alpha q$$

The αq picks up a "valence" element of policy.

[7] When NOMINATE was developed in 1982-83, computer time, disk space, and memory were scarce resources. Howard Rosenthal and I opted for the logit rather than the normal distribution because logit-based probabilities are *formulas* and were faster to

compute! This is no longer an important issue (see Chapter 5 for the practicalities of scaling program design).

[8] This could probably be done by adding another step to the NOMINATE framework similar to the design of QN. Namely, with all other parameters held fixed, estimate the p σ_i's as a separate step.

[9] The structure of NOMINATE was inspired by the work of the psychometricians who developed non-metric multidimensional scaling (Shepard, 1962a,b; Kruskal, 1964a,b) and individual differences scaling methods (Carroll and Chang, 1970). These scaling methods all used an alternating structure to estimate their models.

[10] In early 1983 even with the logit model it took several hours on a state-of-the-art DEC VAX to analyze a single Senate roll call matrix. To analyze a U. S. House matrix with about 500 roll calls required that we run the program overnight in the batch queue. This took at least 7-9 hours of CPU time.

[11] VECTOR FORTRAN allowed the declaration of *bit vectors*. Because a legislator's vote on a particular roll call is either Yea, Nay, or abstain, the legislator's choice required only *two bits* of memory storage on the Cyber 205. The appropriate entry in one bit vector was set to TRUE if the legislator voted Yea. Similarly, the appropriate entry in a second bit vector was set to TRUE if the legislator voted Nay. A logical comparison of the two vectors was used to find abstention. Because bit vectors permitted storing 64 individual roll call votes in just two words of memory, the large memory of the Cyber 205 could be used very efficiently.

[12] Mel Hinich suggested this idea to us in 1986. See Hinich and Roll (1981) for an example of their use in a financial economics model.

[13] As I have shown in Chapter 2 with the cutting plane procedure, the *probit coefficients are the normal vector to the plane that divides the choices*. As the error goes to zero, these coefficients are identified up to a scalar constant and a slight "wiggle" depending on the number of observations (this problem is known as *complete separation* (Silvapulle, 1981; Albert and Anderson, 1984). Although in this instance it is not possible to do inference – there is no error, after all – it also implies that classical inference is technically inappropriate because *before you run the probit you know you will have to throw away the results if you get perfect classification*! Hence the tests *conditional on not getting perfect classification* may not be strictly kosher. I thank Howard Rosenthal for pointing this out to me.

[14] Technically, a slightly stronger statement is true; namely, every parameter is at a maximum conditioned on all the parameters *of the other sets* being held fixed. Each legislator is estimated independent of the remaining legislators and each roll call is estimated independent of the other roll calls.

[15] The consistency issue has not been fully sorted out. It might be that the triple-asympotic conditions of Haberman (1977) can be relaxed. See Rivers (2004) for a discussion of identification and consistency issues in multidimensional spatial voting models.

[16] Computing resources were a bit more limited when we did this experiment in 1993 and 1994!

[17] See Hitchcock (2003) for a short history of MCMC simulation.

[18] Conditioning on the latent utilities makes other types of extensions to the model easy to implement. Ordinal (or even multinomial) responses, mixed variables, and so on, can be dealt with efficiently within this framework (Quinn, 2004).

[19] See Gelman, Carlin, Stern, and Rubin (2000, pp. 325-326) and Gill (2002, pp. 306-311) for a discussion of these conditions. For example, one condition is *irreducibility*. This condition requires that the random walk have a positive probability of jumping from any point to any other point in the distribution.

[20] Software to estimate the IRT model on roll call votes has been developed by Simon Jackman (IDEAL) and Andrew Martin and Kevin Quinn (MCMCpack). On the website for this book

http://pooleandrosenthal.com/Spatial_Models_of_Parliamentary_Voting.htm

under the Chapter 4 link are links to their websites and examples of the output of both software packages.

[21] The output is from IDEAL (Clinton, Jackman, and Rivers, 2004; Lewis and Poole, 2004, p. 117, Figure 5).

[22] Technically these are ±1.96 times the posterior standard deviations centered at the posterior means.

[23] This section draws heavily on the discussion in Lewis and Poole (2004, pp. 109-110).

[24] Jeff Gill (2002) has a comprehensive discussion of the history of this controversy in his excellent textbook on Bayesian methods for social scientists.

View publication stats