Estimating the Distribution of Voter Preferences Using Partially Aggregated Voting Data

James M. Snyder, Jr.
Department of Political Science and Department of Economics
Massachusetts Institute of Technology

February, 2005

Partially aggregated voting returns, especially voting on ballot intitatives and referendums, are a potentially valuable source of data for identifying patterns in voter preferences and for studying questions of political representation. Deacon and Shapiro (1975), Kuklinski (1978), Snyder (1996), Kahn and Matsusaka (1997), and Lewis (1998) exploited such data in earlier work, and Ansolabehere et al (2000, 2002), and others are using similar data currently. In fact, it is arguable that aggregated data are even more relevant than individual level data for studying representation issues. Stimson (1991) states the argument clearly: "For a politician to pay attention to individual views is to miss the main game... The politician must, as a matter of image, appear concerned about individuals, but aggregate opinion is what matters (p. 12)." This point is especially important because aggregate data often exhibit starkly different patterns than individual level data. Aggregate opinion appears to be much more stable than individual opinion, and more predictable as well (Stimson, 1991; Page and Shapiro, 1992) It also appears to be more ideological, at least as measured by Converse's concept of "constraint" (Kuklinski, 1978; Snyder, 1996).

Scholars have developed a variety of empirical models for studying individual level voting data and survey data, which are well-grounded in a decision-theoretic framework (e.g., Poole and Rosenthal, 1997). Currently, however, we lack similar models suited for analyzing partially aggregated voting data. This note begins to fill the gap.

If voting data are aggregated, by legislative districts for example, then it is only possible to recover information about summary measures of voter preferences, such as district means and variances. Also, some assumption must be made about the general form of the within-district distribution of voter preferences, in addition to assumptions about voter behavior. The model below makes the following assumptions about voter behavior. The model below makes the following assumptions: (i) each proposition is viewed as two points in K-dimensional issue space, a Yea alternative and a Nay alternative; (ii) voters have Euclidean preferences, so each voter is characterized by an ideal point and prefers policies closer to this ideal point; (iii) voters are uncertain about the true location of alternatives on each proposition, and this uncertainty can be viewed as random noise added to voters' utilities; (iv) voter ideal

points are normally distributed within each district. Assumptions (i)-(iii) are standard in the theoretical and applied literature on probabilistic voting, and are similar to assumptions in Poole and Rosenthal (1997), Heckman and Snyder (1997), Clinton, Jackman and Rivers (2004), and other work. Assumption (iv) is the main addition, and allows the application to aggregated data.

The most important result proved below is that under assumptions (i)-(iv) the appropriate model to fit aggregated vote data is a simple *linear* factor model (Proposition 2 and Corollary 2). Treating the propositions as variables, the factor loadings describe the propositions, and the factor scores describe the means and variances of the distribution of ideal points in each district. Ordinary principle components analysis of factor analysis may be used to estimate the dimensionality of the issue space, and to estimate linear combinations of the ideal point means.

Three features of the model deserve mention. First, the model allows voter uncertainty to vary across propositions. This is important because the general level of voter knowledge varies widely across propositions, and in most cases is probably more important than variation in the level of knowledge across voters for any given proposition. Few voters in California were unaware that Proposition 13 on the June 1978 ballot would cut property taxes, or that Proposition 10 on the November 1980 ballot, entitled "Smoking and Non-smoking Sections," required restaurants to establish smoking and non-smoking sections. On the other hand, most voters probably knew little about the key issues surrounding Proposition 10 on the 1982 ballot, which allowed counties to merge their superior, municipal and justice courts. Second, if all districts are approximately equally heterogeneous (*i.e.*, the ideal point distributions have the same variance), then there will be exactly as many factors as issue dimensions. If the districts are not equally heterogeneous, then there will $K+1$ factors when the issue space has K dimensions. In this case, K of the factors describe the means of the districts' ideal point distributions scaled by the variances, and the remaining factor describes the variances of these distributions. Third, the distribution of voter ideal points may include dimensions on which all voters in a district have the same ideal point (Proposition

3). This might be try for sectionally defined dimensions, such as "north vs. south." Also, the "quality" of each proposition can be treated as a special case of this type of dimension, which *all* voters have the same ideal point (all voters want higher quality). Sectional issues simply enter as additional linear factors, and quality enters as a constant. Voter "moods" (Stimson, 1991) can also be captured simply as an extra quality dimension.

Finally, I must mention the main limitation of the model. The model assumes that the distribution of voter preferences within each district is symmetric and normal. The assumption of normality can be relaxed (Remark 1 below), but symmetry is necessary to keep the problem simple. Thus, if the actual within-district distributions of voter ideal points are skewed, or if the distributions are unimodal along one dimension but bimodal along another, then the linear factor model is only an approximation of the true model. More work needs to be done to see how adequate this approximation is in practice.

The formal presentation of the model is as follows. Let I be a set of regions, let J be a set of ballot propositions, and let y_{ij} denote the fraction of voters in region i who vote Yea on proposition j. I begin with a basic model, then consider various extension. The basic model consists of the following assumptions:

(A.1) Each proposition j can be described by two points in $\Re^K$, a Yea alternative $\mathbf{x}_j$ and a Nay alternative $\mathbf{s}_j$.

(A.2) All voters have Euclidean preferences. Thus, the utility of a voter with ideal point at $\mathbf{z}$ can be described by $u(\mathbf{z}, \mathbf{x}) = -(\mathbf{x}-\mathbf{z})'(\mathbf{x}-\mathbf{z})$.

(A.3) Voters vote for their most preferred alternative on each proposition.

(A.4) In each region i, the distribution of ideal points is a spherical multivariate normal with mean $\mathbf{z}_i$ and variance σ_i^2. A larger value of σ_i^2 means that voter preferences in region i are more heterogeneous.

Assumptions (A.1)-(A.4) imply that voting on ballot propositions is described by a *linear* factor model, as shown by the following proposition and corollary.

Proposition 1. Assume $(A.1)$-$(A.4)$, let $\mathbf{b}_j = (\mathbf{x}_j - \mathbf{s}_j)/||\mathbf{x}_j - \mathbf{s}_j||$ and let $c_j = (\mathbf{x}_j'\mathbf{x}_j - \mathbf{s}_j'\mathbf{s}_j)/2||\mathbf{x}_j - \mathbf{s}_j||$. Then $y_{ij} = \Phi((\mathbf{z}_i'\mathbf{b}_j - c_j)/\sigma_j)$ for all i and j, where Φ is the standard normal cumulative distribution function.

Proof. By $(A.1)$-$(A.3)$, the set of voters who vote Yea on proposition j is the half-space $Y_j = \{\mathbf{z} \mid \mathbf{z}'\mathbf{b}_j > c_j\}$. By $(A.4)$, the ideal point distribution in region i is given by the joint density $f_i(\mathbf{z}) = (2\pi\sigma_i^2)^{-K/2} exp[-(\mathbf{z} - \mathbf{z}_i)'(\mathbf{z} - \mathbf{z}_i)/2\sigma_i^2]$.

Thus, $y_{ij} = \int ... \int_{Y_j} f_i(\mathbf{z}) d\mathbf{z}$. Letting $\mathbf{v} = (\mathbf{z} - \mathbf{z}_i)/\sigma_i$ and changing variables, $y_{ij} = \int ... \int_{Y_{ij}} (2\pi)^{-K/2} exp[-\mathbf{v}'\mathbf{v}/2] d\mathbf{v}$, where $Y_{ij} = \{\mathbf{v} \mid \mathbf{v}'\mathbf{b}_j > (c_j - \mathbf{z}_i'\mathbf{b}_j/\sigma_i\}$. The integrand is now a multivariate standard normal density, so it is easily integrated (*e.g.*, Anderson, 1958) yielding $y_{ij} = 1 - \Phi((c_j - \mathbf{z}_i'\mathbf{b}_j)/\sigma_i) = \Phi((\mathbf{z}_i'\mathbf{b}_j - c_j)/\sigma_i)$. Q.E.D.

Corollary 1. Let $w_{ij} \equiv \Phi^{-1}(y_{ij})$ for all i and j. Then $\mathbf{w}_j = \mathbf{F}\mathbf{g}_j$ for all j, where $\mathbf{w}_j \equiv (w_{1j} \equiv (w_{1j}, ..., w_{Ij})', \mathbf{g}_j \equiv (b_{j1}, ..., b_{jK}, c_j)'$, and $\mathbf{F} = (\mathbf{f}_1, ..., \mathbf{f}_{K+1}) = ((z_{11}/\sigma_1, ..., z_{I1}/\sigma_I)', ..., (z_{K1}/\sigma_1, ..., z_{KI}/\sigma_I)', (1/\sigma_1, ..., 1/\sigma_I)')$.

Remark 1. Thus, inverting the vote shares y_{ij} using Φ, the resulting variables $\mathbf{w}_1, ..., \mathbf{w}_J$ are *linear* functions of the factors $\mathbf{f}_1, ..., \mathbf{f}_{K+1}$. If σ_i is the same for all i, then there are K factors when the issue space as K dimensions (the $1/\sigma_i$ "factor" is a constant). If the σ_i vary, then there are $K + 1$ factors. Also, results similar to Proposition 1 and Corollary 1 hold for *any* spherical distributions of voter ideal points, not only for normal distributions (see Fang, Kotz and Ng, 1990).

In the basic model voters have perfect information about the propositions and make no "mistakes" when voting. I now assume voters have limited information, modeled as independent, normal, random noise added to voters' utilities. Replace assumption $(A.3)$ above with the following assumption:

$(A.3)'$ The probability that a voter with ideal point $\mathbf{z}$ votes Yea on proposition j is $G_j(u(\mathbf{z}, \mathbf{x}_j) - u(\mathbf{z}, \mathbf{s}_j))$, where G_j is the cumulative distribution function of a normal random variable with mean 0 and variance θ_j^2.

Then we have the following proposition.

Proposition 2. Assume $(A.1)$,$(A.2)$,$(A.3)'$,$(A.4)$, and let $\mathbf{b}_j = 2(\mathbf{x}_j - \mathbf{s}_j)/\theta_j$, $c_j = (\mathbf{x}_j'\mathbf{x}_j - \mathbf{s}_j'\mathbf{s}_j)/\theta_j$, and $\psi_j^2 = \mathbf{b}_j'\mathbf{b}_j$. Then $y_{ij} = \Phi((\mathbf{z}_i'\mathbf{b}_j - c_j)/(\sigma_i^2\psi_j^2+1)^{1/2}))$ for all i and j, where Φ is the standard normal cumulative distribution function.

Proof. By $(A.1)$,$(A.2)$, and $(A.3)'$, the set of voters who vote Yea on proposition j is $Y_j = \{(\mathbf{z},\epsilon) \mid 2\mathbf{z}'(\mathbf{x}_j - \mathbf{s}_j) - \mathbf{x}_j'\mathbf{x}_j + \mathbf{s}_j'\mathbf{s}_j > \epsilon\}$, where $\epsilon \sim N(0,\theta_j^2)$. By $(A.4)$, the distribution of ideal points in region i is given by the density $f_i(\mathbf{z}) = (2\pi\sigma_i^2)^{-K/2} exp[-(\mathbf{z}-\mathbf{z}_i)'(\mathbf{z}-\mathbf{z}_i)/2\sigma_i^2]$. Thus, $y_{ij} = \int ... \int_{Y_j} g_j(\epsilon) f_i(\mathbf{z}) d\epsilon d\mathbf{z}$, where $g_j \equiv G_j'$ is the density function associated with G_j. Substituting $\eta = \epsilon/\theta_j$ and $\mathbf{v} = (\mathbf{z}-\mathbf{z}_i)/\sigma_i$ yields $y_{ij} = \int ... \int_{Y_{ij}} (2\pi)^{-(K+1)/2} exp[-(\mathbf{v}'\mathbf{v} + \eta^2)/2] d\eta d\mathbf{v}$, where $Y_{ij} = \{(\mathbf{v},\eta) \mid \sigma_i\mathbf{v}'\mathbf{b}_j + \mathbf{z}_i'\mathbf{b}_j - c_j > \eta\} = \{(\mathbf{v},\eta) \mid ((\sigma_i\mathbf{v}'\mathbf{b}_j - \eta)/(\sigma_i^2\psi_j^2 + 1)^{1/2}) > ((c_j - \mathbf{z}_i'\mathbf{b}_j)/(\sigma_i^2\psi_j^2 + 1)^{1/2})\}$. Since the integrand is now a multivariate standard normal density it is easily integrated, yielding $y_{ij} = \Phi((\mathbf{z}'_i\mathbf{b}_j - c_j)/(\sigma_i^2\psi_j^2 + 1)^{1/2})$. Q.E.D.

Corollary 2. Let $w_{ij} \equiv \Phi^{-1}(y_{ij})$ for all i and j, and let $\mathbf{w}_j \equiv (w_{1j}, ..., w_{Ij})'$.

(i) If $\sigma_i = \sigma$ for all i, then $\mathbf{w}_j$ is a linear function of the K factors $(z_{11}, ..., z_{I1})', ..., (z_{K1}, ..., z_{KI})'$.

(ii) If $\psi_j^2 = \psi^2$ for all j, then $\mathbf{w}_j$ is a linear function of the $K+1$ factors $(z_{11}/\tilde{\sigma}_1, ..., z_{I1}/\tilde{\sigma}_I)', ...,$ $(z_{K1}/\tilde{\sigma}_1, ..., z_{KI}/\tilde{\sigma}_I)'$, $(1/\tilde{\sigma}_1, ..., 1/\tilde{\sigma}_I)'$, where $\tilde{\sigma}_i = (\sigma_i^2\psi^2 + 1)^{1/2}$.

(iii) If σ_i^2 and ψ_j^2 are "large", then $\mathbf{w}_j$ is approximately a linear function of the $K+1$ factors $(z_{11}/\sigma_1, ..., z_{I1}\sigma_I)', ...,$ $(z_{K1}/\sigma_1, ..., z_{KI}/\sigma_I)'$, $(1/\sigma_1, ..., 1/\sigma_I)'$.

Proof. Parts (i) and (ii) are obvious. The proof of (iii) follows by noting that if σ_i^2 and ψ_j are large, then $(\sigma_i^2\psi_j^2 + 1)^{1/2} \approx \sigma_i\psi_j$.

Remark 2. To see what "large" means in Corollary 2, note that ψ_j^2 measures how informed voters are about proposition j – larger values of ψ_j^2 mean fewer voter mistakes. Normalize by setting $(\mathbf{x}_j - \mathbf{s}_j)'(\mathbf{x}_j - \mathbf{s}_j) = 1$, and suppose voters with $\mathbf{z} = \mathbf{x}_j$ vote for $\mathbf{x}_j$ at least 85% of

the time (other voters will make mistakes more often). Then $\psi_j^2 = 4/\theta_j^2 \geq 4\Phi^{-1}(.85) \approx 4.3$. Assuming that preference heterogeneity is at least as important a factor as voter information, $\sigma_i^2\psi_j^2 \geq 18.5$, and the discrepancy between $(\sigma_i^2\psi_j^2+1)^{1/2}$ and $\sigma_i\psi_j$ is only 2.5% or less.

The basic model above assumes that voters' preferences within each region vary across all of the K dimensions. I now extend the model to incorporate issues on which all voters within a given region have the *same* ideal point. This might be true for issues dealing with the geographic distribution of resources. Also, the "quality" of a proposition can be treated as a dimension on which all voters have the same ideal point; all voters want higher quality.

$(A.4)'$ In each region i, the distribution of ideal points with respect to dimensions $1, ..., K-1$ is a spherical multivariate normal with mean $\hat{\mathbf{z}}_i$ and variance σ_i. With respect to dimension K, either all voters have $z_{iK} = 1$, or all voters have $z_{iK} = 0$.

Proposition 3. Assume $(A.1), (A.2), (A.3), (A.4)'$, and let $\hat{\mathbf{x}}_j \equiv (x_{j1}, ..., x_{j,K-1})$, $\hat{\mathbf{s}} \equiv (s_{j1}, ..., s_{j,K-1})$, $\mathbf{b}_j = ((\mathbf{x}_j - \mathbf{s}_j)/||\hat{\mathbf{x}}_j - \hat{\mathbf{s}}_j||)$ and $c_j = ((\mathbf{x}_j'\mathbf{x}_j - \mathbf{s}_j'\mathbf{s}_j)/(2||\hat{\mathbf{x}}_j - \hat{\mathbf{s}}_j||))$. Then $y_{ij} = \Phi((\mathbf{z}_i'\mathbf{b}_j - c_j)/\sigma_1)$ for all i and j, where Φ is the standard normal cumulative distribution function.

Proof. By $(A.1)$-$(A.3)$, the set of Yea voters on j is $Y_j = \{\mathbf{z} \mid \mathbf{z}'\mathbf{b}_j > c_j\}$. Thus, letting $\hat{\mathbf{z}} \equiv (z_1, ..., z_{K-1})$ and $\hat{\mathbf{b}}_j \equiv (b_{j1}, ..., b_{j,K-1})$, for each region i with $z_{iK} = 1$ the set of Yea voters is $Y_{ij} = \{\mathbf{z} \mid \hat{\mathbf{z}}'\hat{\mathbf{b}}_j > c_j - b_{jK}\}$. By $(A.4)'$, the ideal point density in region i is: $f_i(\mathbf{z}) = (2\pi\sigma_i^2)^{-(K-1)/2} exp[-(\hat{\mathbf{z}} - \hat{\mathbf{z}})'(\hat{\mathbf{z}} - \hat{\mathbf{z}})/2\sigma_i^2]$ if $z_K = 1$, and $f_i(\mathbf{z}) = 0$ if $z_K \neq 1$. Thus, $y_{ij} = \int ... \int_{Y_{ij}} f_i(\mathbf{z}) d\mathbf{z}$, or substituting $\hat{\mathbf{z}} \equiv (\hat{z_1}, ..., \hat{z_{K-1}})$ and $\mathbf{v} = (\hat{\mathbf{z}} - \hat{\mathbf{z}}_i)/\sigma_i, y_{ij} = \int ... \int_{Y_{ij}'} (2\pi)^{-K/2} exp[-\mathbf{v}'\mathbf{v}/2] d\mathbf{v}$, where $Y_{ij}' = \{\mathbf{v} \mid \mathbf{v}'\hat{\mathbf{b}}_j > (c_j - b_{jK} - \hat{\mathbf{z}}_i'\hat{\mathbf{b}}_j)/\sigma_i\}$. Integrating, $y_{ij} = \Phi((\hat{\mathbf{z}}_i'\hat{\mathbf{b}}_j + b_{jK} - c_j)/\sigma_i) = \Phi((\mathbf{z}_i'\mathbf{b}_j - c_j)/\sigma_i)$. Similarly, for each region i with $z_{iK} = 0$ the set of Yea voters is $Y_{ij} = \{\mathbf{z} \mid \hat{\mathbf{z}}'\hat{\mathbf{b}}_j > c_j\}$ and the distribution of ideal points is: $f_i(\mathbf{z}) = (2\pi\sigma_i^2)^{-(K-1)/2} exp[-(\hat{\mathbf{z}} - \hat{\mathbf{z}})'(\hat{\mathbf{z}} - \hat{\mathbf{z}})/2\sigma_i^2]$ if $z_K = 0$, and $f_i(\mathbf{z}) = 0$ if $z_K \neq 0$. Changing variables and integrating yields $y_{ij} = \Phi((\hat{\mathbf{z}}_i'\hat{\mathbf{b}}_j - c_j)/\sigma_i) = \Phi((\mathbf{z}_i'\mathbf{b}_j - c_j)/\sigma_i)$. Q.E.D.

Remark 3. Inverting the y_{ij} using Φ yields a linear factor model, just as in corollary 1.

Also, Proposition 3 and its corollary are easily generalized to "geographic" issues that have more than two values, and to issue spaces with more than one such issue. Finally, combining "geographic" issues and limited information yields results analogous to Proposition 2 and its corollary.

Remark 4. To see how quality can be treated as a dimension on which all voters have the same ideal point, let the Kth dimension represent quality, and label the dimension's axis so that a higher value means *lower* quality. The preferences of a voter whose ideal point along dimensions $1, ..., K-1$ is at $\hat{\mathbf{z}} \equiv (z_1, ..., z_{K-1})$ can then be represented by $u(\mathbf{z}, \mathbf{x}) = -(\hat{\mathbf{x}}-\hat{\mathbf{z}})'(\hat{\mathbf{x}}-\hat{\mathbf{z}}) - x_K^2$, where $\hat{\mathbf{x}} \equiv (x_1, ..., x_{K-1})$. Letting $z_K = 0, u(\mathbf{z}, \mathbf{x}) = -(\mathbf{x}-\mathbf{z})'(\mathbf{x}-\mathbf{z})$. That is, it is as if each voter has Euclidean preferences in a K-dimensional space, with ideal point along the Kth dimension at zero. The quality dimension does not enter as a separate factor, however, since all regions have the same mean ideal point along this dimension.

References

Anderson, T.W. 1958. *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley and Sons.

Ansolabehere, Stephen, James M. Snyder, Jr., and Jonathan Woon. 2000. "Why Did a Majority of Californians Vote to Limit Their Own Power?" Unpublished manuscript, Massachusetts Institute of Technology.

Ansolabehere, Stephen, James M. Snyder, Jr., and Ruimin He. 2002. "Evidence of Virtual Representation: Reapportionment in California." Unpublished manuscript, Massachusetts Institute of Technology.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98: 355-370.

Deacon, Robert, and Perry Shapiro. 1975. "Private Preference for Collective Goods Revealed through Voting on Referenda." *American Economic Review* 65: 943-955.

Fang, K.T., S. Kotz and K.W. Ng. 1990. *Symmetric Multivariate and Related Distributions*. New York: Chapman and Hall.

Heckman, James J., and James M. Snyder, Jr. 1997. "Linear Probability Models of the Demand for Attributes with an Empirical Application to Estimating the Preferences of Legislators." *RAND Journal of Economics* 28:S142-189.

Kahn, Matthew E., and John G. Matsusaka. 1997. "Demand for Environmental Goods: Evidence from Voting patterns on California Initiatives." *Journal of Law and Economics* 40: 137-173.f

Kuklinski, James H. 1978. "Representativeness and Elections: A Policy Analysis." *American Political Science Review* 72: 165-177.

Lewis, Jeffrey B. 1998. *Who Do Representatives Represent?* Unpublished dissertation, Massachusetts Institute of Technology.

Page, Benjamin I., and Robert Y. Shapiro. 1992. *The Rational Public*. Chicago: University of Chicago Press.

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

Snyder, James M., Jr. 1996. "Constituency Preferences: California Ballot Propositions, 1974-90." *Legislative Studies Quarterly* 21: 463-488

Stimson, James A. 1991. *Public Opinion in America: Moods, Cycles, and Swings*. Boulder, CO: Westview Press, 1991.