Technical Appendix to Estimating Voter Preference Distributions from Individual-level Voting Data

Jeffrey B. Lewis *
jblewis@polisci.ulca.edu.
Department of Political Science
UCLA

August 24, 2001

This companion appendix to "Estimating Voter Preference Distributions from Individual-level Voting Data" (*Political Analysis* 9(3):275–297, hereafter Lewis 2001) provides: a more detailed derivation of the estimation technique and its implementation (Section A), additional Monte Carlo results (Section B), and a description of how the data sets used in Section 5 of Lewis (2001) were constructed (Sections C and D).

A A fuller description of the ideal point distribution estimation

The model presented in Lewis (2001) allows the estimation of the mean and variance of the distribution of voter ideal points along a single dimension across any number of predetermined voter groups. If the number of voter groups is held constant, the estimated group means and variances can be consistently estimated as the number of voters grows large. The model is derived in the usual way from a simple spatial model in which voters have quadratic preferences. By parameterizing not only the mean but also the variance of the ideal point distribution across groups, the estimator below extends a general class of two-parameter Rasch item response models (see Bock & Aitkin, 1981) that have been adapted to legislative settings by Bailey (1998) and Londregan (2000). Because the data sets to which the method is applied to are quite large special attention is

*Address: Political Science Department, University of California Los Angeles, Bunche Hall, Los Angeles CA 90095-1572.

Index, parameter, and variable definitions

Quantity/index	*Symbol*
Voters	$l = 1, 2, \ldots, N$
Items (Propositions)	$k = 1, 2, \ldots, K$
Groups	$g = 1, 2, \ldots, G$
Vote patterns	$i = 1, 2, \ldots, I$
	$q = 1, 2, \ldots, Q$
Quadrature points	
Item intercepts	$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$
Item slopes	$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_K)$
Group means	$\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_G)$
Group standard deviations	$\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_G)$
Votes pattern i in group g	$\mathbf{V}_{ig} = (V_{ig1}, V_{ig2}, \ldots, V_{igK})$
Number of voters in group g casting pattern i	c_{ig}
Ideal point of voter l	θ_l
qth quadrature point	X_q
Estimated number of voters located at quadrature point q voting "yes" on item k	$\bar{y}_{kgq}$
Estimated number of voters located at quadrature point q voting "no" on item k	$\bar{n}_{kgq}$

Table 1: *Model indices and parameters.*

paid to techniques that can be used to reduce the computational burden.

The joint likelihood for the model given in Lewis (2001, p. 281) is

$$L(\mathbf{V}, \boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \prod_l \prod_k \Phi(\alpha_k + \beta_k\theta)^{v_{lk}}(1 - \Phi(\alpha_k + \beta_k\theta))^{(1-v_{lk})}\phi(\theta|\mu_{l(g)}, \sigma_{l(g)})$$

where $g(l)$ is a function returning the group affiliation of voter l. All of the variables and indices are given in Table 1 and in Lewis (2001). Integrating out over the unobserved random variable θ, we find the marginal maximum likelihood,

$$L(\mathbf{V}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \prod_i \int \prod_k \Phi(\alpha_k + \beta_k\theta_l)^{v_{lk}}(1 - \Phi(\alpha_k + \beta_k\theta_l))^{(1-v_{lk})}\phi(\theta|\mu_{g(l}, \sigma_{g(l)})d\theta.$$

The maximization of this likelihood yields consistent estimators of the proposition parameters and the group means and standard deviations. Maximization of the likelihood is not straightforward. In particular, the estimation involves calculation of an integral by numerical methods for each observation. Since the datasets to which this method will be applied are generally very large, these calculations are burdensome. Moreover, because an integral must be approximated, derivative-based maximization techniques can be problematic.

In order to simplify the estimation of the model, I follow Bock & Aitkin (1981) in grouping the data into patterns[1] and in employing an EM algorithm to overcome computational problems related to the direct maximization of the likelihood presented above.

Grouping of the data is the first computational savings that can be achieved. All we know about each voter is the pattern of votes that they cast across the K items and the group to which they belong. Any two voters casting the same pattern of votes and belonging to the same group are interchangeable and thus must make the same contribution to the likelihood. Following this logic, we can group the data by patterns of votes and group membership. Letting i index patterns of votes, define V_{ig} to be ith vote pattern in group g. Similarly define c_{ig} be the number of voters

[1] Londregan 2000 also points out this possibility in a different context.

in group g casting pattern V_{ig}. The likelihood can then be reformed as

$$L(\mathbf{V}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\sigma}) = \prod_i \prod_g \left[\int \prod_k \Phi(\alpha_k + \beta_k\theta)^{V_{igk}} \left[1 - \Phi(\alpha_k + \beta_k\theta)\right]^{(1-V_{igk})} \phi(\theta|\mu_g,\sigma_g)d\theta \right]^{c_{ig}}.$$

similarly, the log likelihood can be written as

$$\ln L(\mathbf{V}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\sigma}) = \sum_i \sum_g c_{ig} \ln \left(\int \prod_k \Phi(\alpha_k + \beta_k\theta)^{V_{igk}} \left[1 - \Phi(\alpha_k + \beta_k\theta)\right]^{(1-V_{igk})} \phi(\theta|\mu_g,\sigma_g)d\theta \right).$$

Even greater numerical efficiencies can be achieved through the application of an EM algorithm (following Bock & Atkin 1981). This algorithm can be thought of as a method for maximizing certain likelihood functions in the presence of missing information (Dempster, Laird & Rubin 1977). The method proceeds in Expectation-steps and Maximization steps. In the E-step, the expectations of the missing information given the data and provisional values of the parameters to be estimated are calculated. In the M-step, the likelihood is maximized substituting the estimated expected values from the E-step for the missing information. The procedure is then repeated with the next E-step using the new estimates of the parameters from the M-step. The process ends when the estimated parameters are stable from one iteration to the next.

To reconceptualize the likelihood developed above in the EM context, it is useful to define a few additional functions. First, the probit function, P_k, describes the probability of observing a particular vote choice, V_{igk}, given the item parameters, group parameters, and the Z-transformed voter ideal point, $Z = \frac{\theta-\mu_g}{\sigma_g}$. This function is written as

$$P_k(\mathbf{V}_{igk}|\alpha_k,\beta_k,\mu_g,\sigma_g,Z) = \Phi(\alpha_k + \beta_k(\mu_g + \sigma_g Z))^{V_{igk}} \left[1 - \Phi(\alpha_k + \beta_k(\mu_g + \sigma_g Z))\right]^{(1-V_{igk})}.$$

The function P describes the probability of observing a particular pattern of votes across the K choice items as

$$P(\mathbf{V}_{ig}|\boldsymbol{\alpha},\boldsymbol{\beta},\mu_g,\sigma_g,Z) = \prod_k P_k(\mathbf{V}_{igk}|\alpha_k,\beta_k,\mu_g,\sigma_g,Z).$$

The full marginal log likelihood can then befined as,

$$\ln L(\mathbf{V}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\sigma}) = \sum_i \sum_g c_{ig} \ln \left(\int P(\mathbf{V}_{ig}|\boldsymbol{\alpha},\boldsymbol{\beta},\mu_g,\sigma_g,Z)\phi(Z)dZ \right)$$

where ϕ is the standard-normal density function. This likelihood cannot be evaluated analytically and must be approximated. This is achieved by replacing the continuous variable Z and density ϕ with the discrete variable X and distribution (or weight) function A. Letting q index the points of support of X and letting Q be the number of such points, we can approximate the log-likelihood as:

$$\ln L(\mathbf{V}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\sigma}) \approx \sum_i \sum_g c_{ig} \ln \left(\sum_{q=1}^{Q} P(\mathbf{V}_{ig}|\boldsymbol{\alpha},\boldsymbol{\beta},\mu_g,\sigma_g,X_q)A(X_q) \right). \tag{1}$$

The accuracy of this approximation is determined by the number of quadrature points Q. As shown in Section B, quite reasonable approximations are achieved with as few as 6 points of quadrature.

Equation 1 could be maximized over the item and group parameters directly. However, this problem turns is much easier to solve by an EM approach. In describing this approach, it is convenient to suppose that the distribution of voters actually follows the distribution of X rather than Z. That is, we will proceed as if there were a certain number of voters located at each quadrature point (point in the support of X) within each group. Voters in each group and at each quadrature point can be partitioned into those voting for and those voting against a certain item. Letting y_{kgq} and n_{kgq} be the number of voters in group g and at point X_q voting for and against item k respectively, we can write the likelihood as a grouped probit,

$$\begin{aligned} \ln L^*(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\sigma}) = \textstyle\sum_k \sum_g \sum_q \quad & A(X_q)(y_{kgq} \ln \Phi\left(\alpha_k + \beta_k(\mu_g + \sigma_g X_q)\right) \\ & + \bar{n}_{kgq} \left[1 - \ln \Phi\left(\alpha_k + \beta_k(\mu_g + \sigma_g X_q)\right)\right]). \end{aligned} \tag{2}$$

If we knew the values of the ys and ns, it would be straightforward to maximize this grouped probit function over the item and group parameters. However, these ys and ns are "missing data." These "missing" values are replaced by their expected values given provisional item and group parameter estimates.

The expected ys and ns are straight forward to calculate. The distribution of X given a particular observed vector of votes ($\mathbf{V}_{ig}$) and set of item and group parameters can be found by Bayes'

rule to be

$$P'(X_q|\mathbf{V}_{ig}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mu_g, \sigma_g) = \frac{P(\mathbf{V}_{ig}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mu_g, \sigma_g, X_q)A(X_q)}{\sum_q P(\mathbf{V}_{ig}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mu_g, \sigma_g, X_q)A(X_q)}.$$

The expected number of voters at a particular point (X_q) and group (g) that vote yes on a given item (k) is found by summing the expected number of voters casting a particular pattern for which $V_{ijk} = 1$ over all such patterns:

$$\mathrm{E}(y_{kgq}) = \sum_i V_{igk} c_{ig} P'(X_q|\mathbf{V}_{ig}, \mu_g, \sigma_g).$$

The number of voters at a particular point voting no on a particular item is similarly determined to be

$$\mathrm{E}(n_{kgq}) = \sum_i (1 - V_{igk}) c_{ig} P'(X_q|\mathbf{V}_{ig}, \mu_g, \sigma_g).$$

The maximization of Equation 2 is the M-step and the estimation of the expected values of the ys and ns is the E-step. The computational advantages of this method comes from the M-step only involving loops over $q \times g$ elements as opposed to the $I \times g$ vote patterns that would be looped over in the direct maximization of the marginal likelihood (Equation 1). The use of the EM algorithm is both faster and more reliable in this context.[2]

One drawback of the EM algorithm in contrast to direct maximization using standard techniques is that there is no straightforward way to estimate the standard errors of the estimates. Rather, once estimates are arrived at, the standard errors must be calculated by using gradient or Hessian-based estimators derived from the full log-likelihood function.[3] For this reason, EM does not allow the analyst to avoid calculation of (at least) the gradient of the full likelihood.

Given parameter estimates *a posteriori* estimates of the ideal point distribution over the set of quadrature points can be written as

$$p(X_q|\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \hat{\mu}_g, \hat{\sigma}_g) = \frac{\bar{y}_{kgq} + \bar{n}_{kgq}}{N}$$

for any $k = (1, 2, \ldots, K)$, The *a posteriori* distribution of ideal points for a particular group can be

[2] In experiments, I have found that the EM approach is several times faster than the direct maximization of the likelihood and more likely to converge.

[3] All standard error estimates presented in Lewis (2001) are based on the outer product of the gradient vector.

approximated by the (discrete) distribution of $\mu_g + \sigma_g X$. In Lewis (2001), this discrete distribution is smoothed using standard kernel density techniques.

B Monte Carlo Results

Given the relatively complex nature of the maximization problem described in the previous Section and the strong assumptions that underlie the model, Monte Carlo experimentation on the efficacy of the method and its robustness to small violations in its assumptions is warranted. In this section, I present the results of extensive Monte Carlo experiments to address these questions. In general, the experiments reveal that the model is effective in recovering inter-group mean and variance differences. Even with as few as several hundred observations, and moderate violations of the distributional or independence assumptions quite reasonable estimates are produced. The following discussion presents the results of four sets of experiments focusing on: the number of voters, the number of quadrature points used in the estimation, violations of the assumed conditional independence of the items, and violations of the assumed normality of the distribution of ideal points within groups.

For all of the experiments, the number of groups is held fixed at five and the observations are assumed to be evenly distributed among the five groups. Except where otherwise noted, the number of voters (N) is held fixed at 2,500. Some of the results need to be interpreted partially in terms of this restriction. In particular, the number of observations required to gain good group parameter estimates would be larger if the number of groups considered were larger and smaller if the number of groups considered were smaller. Other results, such as those related to violations of the assumptions are relatively unaffected by the number of groups (holding the number of observations per group constant).

Table 2 shows the values of all of the parameters used in the experiments. Notice that the slopes include both positive and negative values. That is, in some cases, the more "liberal" voters are more likely to vote for a particular item, and in other instances, they are more likely to vote against. Similarly, some groups have mean positions of less than zero and others have means above zero. Because the distribution of ideal points is only defined up to a linear transformation, I have normalized the space such that group one's mean ideal point is zero and its standard deviation is

Parameter Values Used in the Monte Carlo Experiments

Item Parameters			Group Parameters		
Item	Intercept (α)	Slopes (β)	Group	Mean (μ)	Standard Deviations (σ)
1	0.0	0.7	1	0.0	1.0
2	0.1	0.5	2	0.3	1.0
3	0.2	−0.5	3	1.7	0.8
4	0.3	1.0	4	0.4	1.2
5	0.4	−1.2	5	−0.2	1.4
6	0.5	−0.3			
7	0.3	−0.7			
8	0.8	0.4			
9	0.4	0.8			
10	−0.2	0.3			
11	0.5	0.7			
12	0.1	−0.5			

Table 2: *Shows the values of each of parameters used in the Monte Carlo experiments described and presented below.*

one. Even given this normalization one can arrive at two equivalent solutions to the problem. In one solution, higher ideal points reflect the more conservative positions and in the other solution, lower values reflect the conservative position. In either case, *exactly* the same information and implications are revealed. If we call the parameters shown in Table 2 the "true" values, there is an equivalent set of "reflected" parameters that are observationally equivalent in which the sign on the group means and item slopes are all reversed. In the experiments presented below, if the "reflected" solution was found I transformed the parameters to reflect the "true" solution.

B.1 How many voters must be observed?

A natural question is how many voters must be observed in order to assure stable and reliable parameter estimates. To investigate this question, I performed a series of experiments with five groups and six items in which the number of observations was varied between 400 and 25,600 observations. 500 trials were run for each number of observations.[4] As mentioned above, the

[4]The number of observations was set to 400, 800, 1,600, 3,200, 6,400, 12,800, and 25,600.

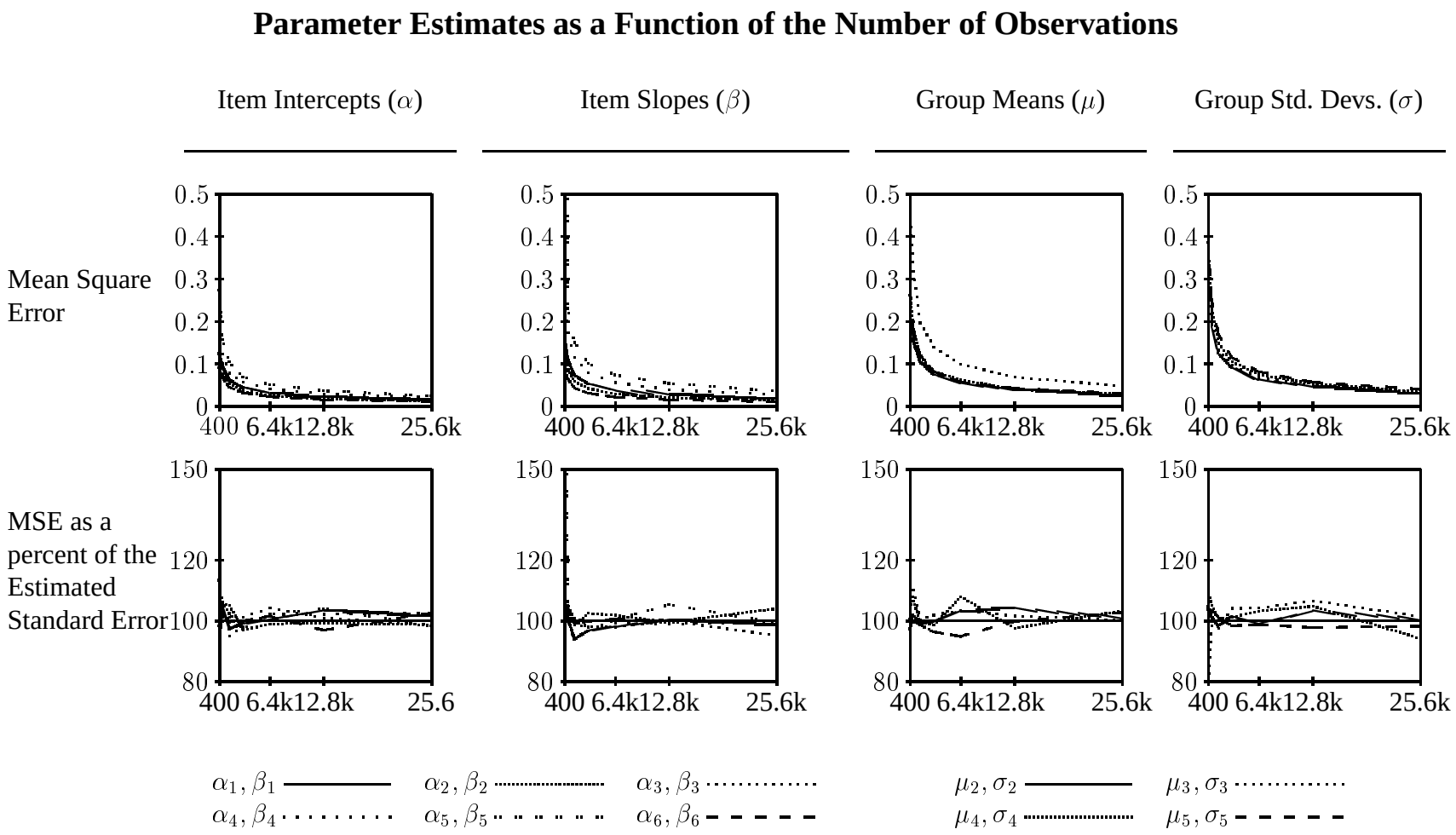

Figure 1: *Shows the results of Monte Carlo experiments in which the total number of observations is varied from 400 up to 25,600 by repeated doublings. The observations are evenly divided among the 5 groups. Ten percent of the total voter-item pairs are randomly assigned to be missing. Overall, the simulations reflect the consistency of the estimators.*

conclusions here are in part dependent upon the number of groups. Particularly with regard to the estimation of the group means and variances, the number of observations per group is more relevant than the total number of observations. In this case, the number of observations per group varies from 80 up to 6,400.

Figure 1 presents the results of these experiments. As in the previous figure, the top row of panels show the observed MSEs for the various model parameters and the lower panels show the average estimated standard errors as a percentage of the observed MSEs. The MSEs fall as the number of observations increases in the usual fashion. Note that with only 400 observations (80 per group), the MSEs are quite large. In fact, with fewer than 400 observations the estimates became extremely unstable and convergence of the maximization routine occasionally failed. As in the previous experiments, the larger βs, μs, and σs were particularly unstable. The standard errors are quite reliable above 800 observations (160 per group). The estimated parameter uncertainty never deviates from the observed uncertainty by as much as 10 percent.

B.2 How many quadrature points to use

As described in Section A, the statistical model requires the approximation of the normal integral. In this set of experiments, I consider how many points of quadrature (elements of the support of the distribution of X) are required to generate stable and reliable estimates. The question is important because there is a tradeoff between increasing the accuracy of the approximation and increasing the computational burden. In their simpler model, Bock & Aitkin (1981) report that as few as 3 or 5 points of quadrature were sufficient for stable parameter estimates. Here I find that the quite stable parameter estimates are achieved with about 6 quadrature points. Little additional advantage is gained by further increasing the number of points.

Figure 2 present the results of these experiments. With two or three points of quadrature the parameters were very unstable. The estimates quickly settle down and by 6 points the estimates are quite stable both in terms of their observed MSEs and the accuracy of their estimated standard errors.

In additional experiments (not presented here), I have found that more quadrature points may be required in situations in which the group means are more widely spread and the slopes are

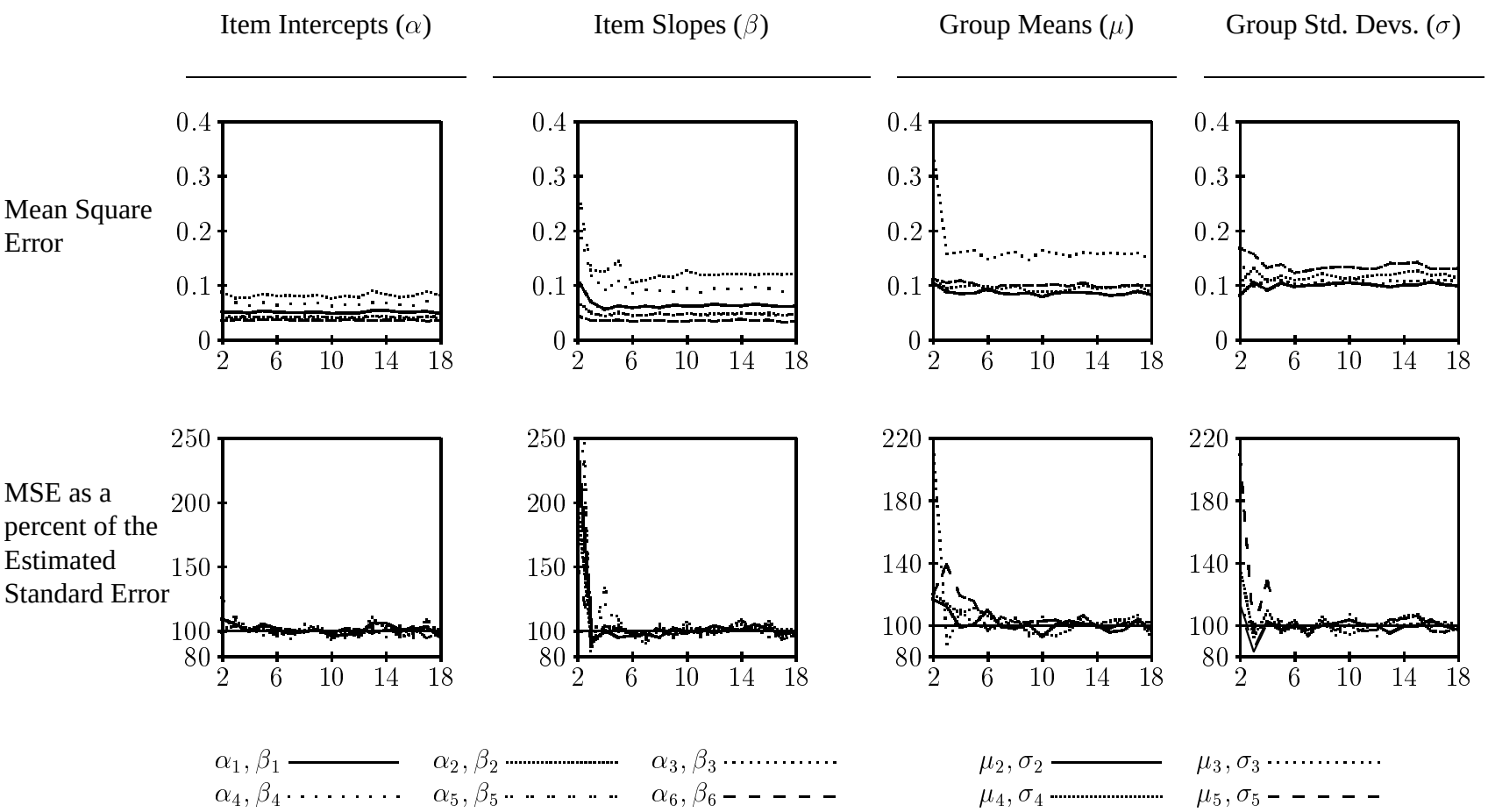

Figure 2: *Shows the results of a series of Monte Carlo experiments in which the number of quadrature points used to approximate the normal integral is varied from 2 to 14. Notice that while the estimates are somewhat unstable when the number of quadrature points is five or fewer, very similar results are obtained for any number of quadrature points from 6 to 18.*

larger in absolute value. In those situations, the exact shapes of the tails of the distribution are more important for the estimation, and additional quadrature points are required. Twelve points of quadrature seem to be sufficient for all parameter configurations that I have considered.

B.3 What happens if the items are not independent?

The next two sets of experiments, I consider the robustness of the model to violations of two of its assumptions. In this section, I will consider the effect of residual correlation between two of the items. This sort of dependence might occur in applications where two items are either similar in content or (in the case of propositions) backed by the same group. In such situations, the observed responses might be correlated beyond what is predicted by the model. On the other hand, we might also observe votes on propositions on the same issues on the same ballot (for example, insurance reform in California) where voting for one reduces the likelihood of voting for the other *ceteris paribus*.

Figure 3 presents the results of a set of experiments in which the residual correlation between items 1 and 2 is manipulated. That is, additional dependence between votes on items 1 and 2 beyond that implied by the model is introduced into the data. Unlike the previous figures, in this figure, the concern is with the bias of the estimates. Rather than report the MSEs of the estimates, the top panels of the figure display the average observed estimates as a percentage of the true parameter values. Percentages greater than 100 represent coefficients that are inflated relative to their true values, and percentages less than 100 reflect parameters that are attenuated relative to their true values.[5]

As the residual correlation between items 1 and 2 increases, we see that the estimated β_1 and β_2 are inflated and the estimated α_1 and α_2 are attenuated as the model tries to accommodate the additional observed correlation between items 1 and 2. However, effects on the other slopes and intercepts are modest. More importantly, the figure reveals minimal effects on the group means and variances. The estimated group means are slightly attenuated, meaning that the inter-group differences are understated in this situation. Even with a very high residual correlation between items 1 and 2, the degree of attenuation of the estimated group means is less than 10 percent. The

[5]For α_1, which is 0 by definition, this ratio cannot be calculated.

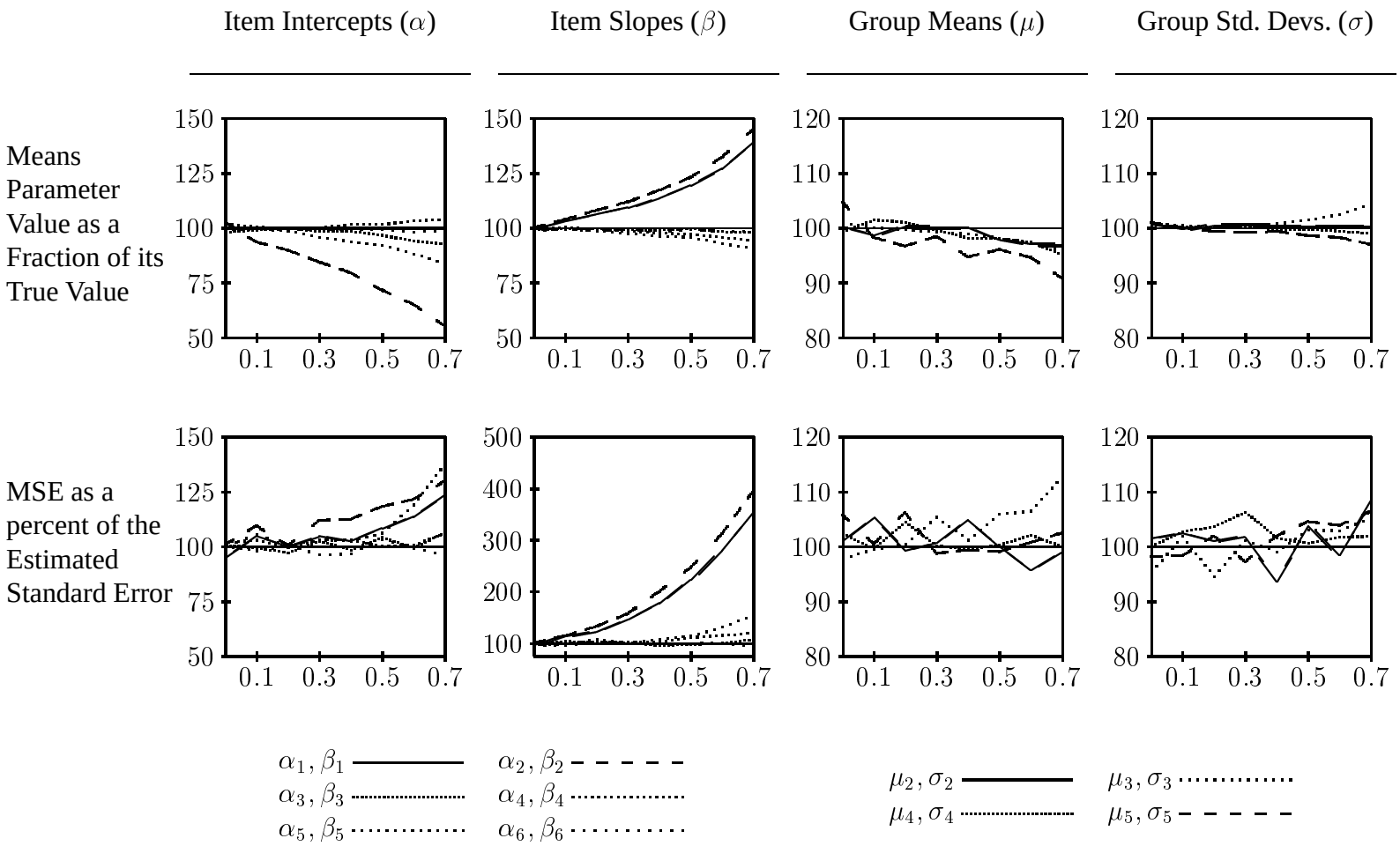

Figure 3: *In these simulations the residual correlations of the propensities to vote for items 1 and 2 are varied between 0, the assumed degree of residual correlation, and* 0.7. *The figure shows that as the correlation between these error terms increases, estimates of the corresponding* β*'s are inflated (in absolute value), the group means (*μ*s) are slightly deflated while the group variances (*σ*s) are largely unaffected. The estimated standard errors greatly understate the true uncertainty (MSEs) of the item parameters (particularly those that are badly biased—*β_1 *and* β_2*). Overall, non-independence of a pair of items even at high levels, only modestly biases the groups parameters and their standard errors.*

observed effects on the estimated variances is similarly minimal.

Turning to the lower panels, we see that the estimated standard errors deviate greatly from the observed MSEs for the intercepts and slopes associated with items 1 and 2. This is mainly due to the bias in the estimates introduced by the residual correlation. The quality of the standard error estimates of the other intercepts is also effected, but in no case are the standard errors of the other slopes and intercepts overstated or understated by more than 50 percent. The standard error estimates for the means and variances are quite accurate throughout.

This suggests that while violations to the assumption of conditional item independence are problematic for the estimation of the item parameters, the group parameters are quite robust to this violation. Even with large residual correlations between two items, the estimated group parameters can be ascertained with a quite high degree of accuracy and reliability.

B.4 Non-normal within-group ideal point distributions

In another set of experiments, I consider the robustness of the model to moderate violations of the assumption that within-group voter distributions are normal. There are many ways in which this could be tested. Here, within-group ideal point distributions continue to be symmetric, unimodal, and have finite variance. Deviations from normality are created by drawing the ideal-point distribution from mixtures of uniform and normal distributions having the same mean and variance. The larger the share of the mixture that comes from the uniform distribution the greater the deviation from normality.[6]

Figure 4 presents the results of Monte Carlo experiments in which the degree of non-normality is varied. The degree of non-normality is indexed by the fraction of the ideal point distribution that is drawn from a uniform rather than a normal distribution. When less than one-half of the ideal points are drawn from the uniform distributions, the effect of non-normality on the estimates are minimal. However, very high levels the non-normality does affect the estimates. In particular, the

[6]More extreme deviations from normality could be considered. However, the assumption of unimodality within well-specified groups does not seem overly strong. Assuming skewed distributions without allowing that skew to vary across groups seems artificial. On the other hand, allowing the skew to vary across groups without modeling that variation is a much more serious specification error than "non-normality." Extending the model in this way is feasible although I have not pursued such an extension. Thus, I only consider unimodal and symmetric violations to the normality assumption.

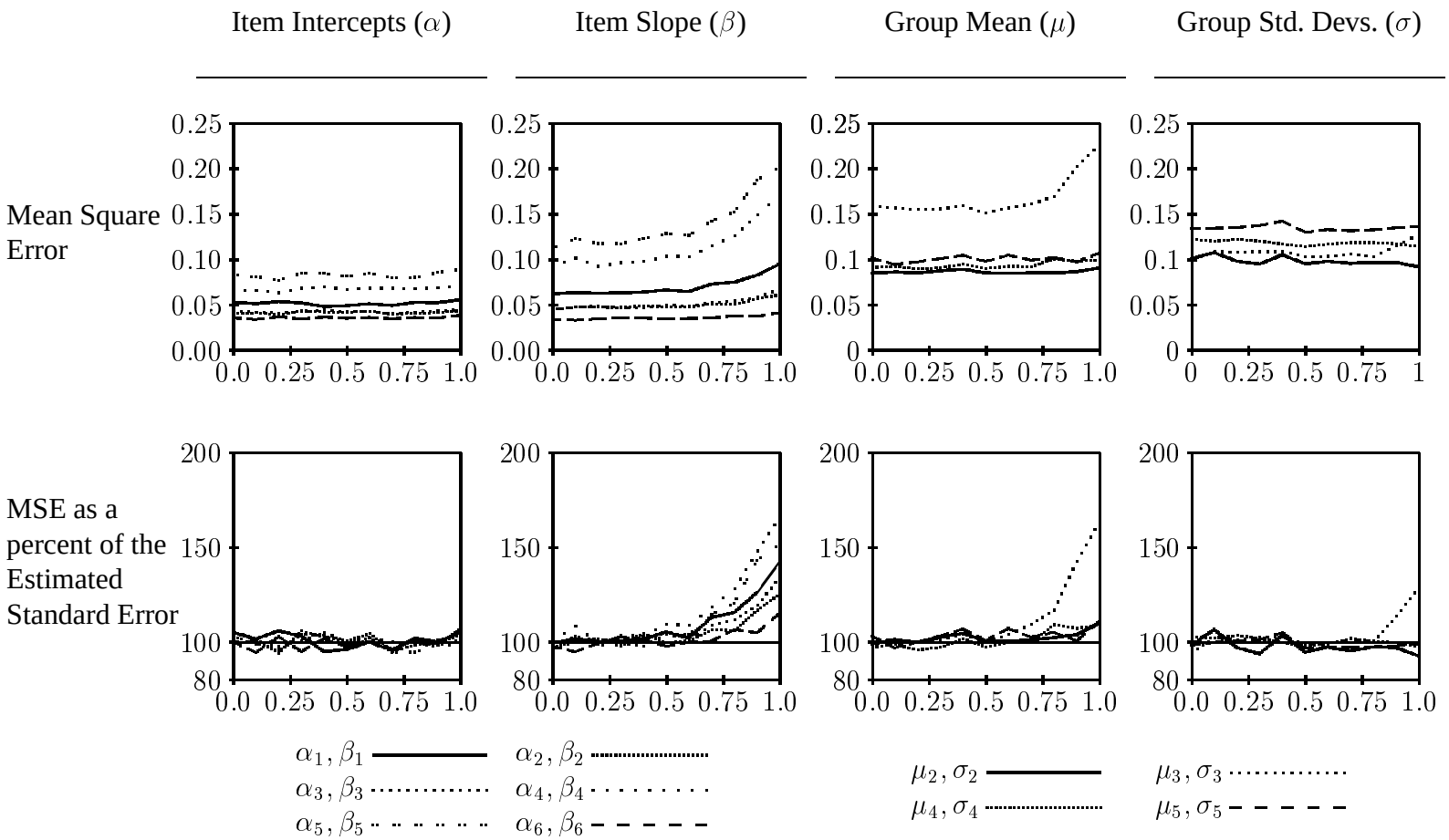

Figure 4: *Presents the results of Monte Carlo experiments in which the the distribution of voter ideal points (θs) in the simulated data is drawn from a mixture of uniform and normal distributions. For each panel the x-axis is the fraction of this mixture that is contributed by the uniform component. Notice that as the mixture distribution becomes increasing uniform, the slopes (βs) and group means (μs) become increasingly biased (MSEs increase and inflate relative to the estimated standard errors).*

item slopes (βs) and group means (μs) are considerably biased when the ideal point distribution is mainly drawn from the uniform distribution. The reason for this is that these parameters are particularly sensitive to the tail probabilities (particularly the extreme groups and lopsided items as noted above) that vary greatly between the normal and uniform distributions.

We can conclude that if the degree of non-normality is relatively small—if the distribution of ideal points is reasonably bell shaped— the estimated parameters will be reliable and the estimated standard errors accurate.

C The Los Angeles County Voter Data

Voting in Los Angeles County uses a punch card system. Each voter punches holes in a computer card corresponding to particular candidate and ballot measure choices. The cards are then read by a computer and the vote totals tallied. As a by-product of this process an "electronic image" (a series of numbers representing the punch pattern) of each cast ballot is made and stored on magnetic tape. These so called ballot image tapes have been made available to the public for several recent election cycles.[7] Some analysis of these data has been undertaken by scholars at Cal Tech (Ken McCue) and UCSD (Elisabeth Gerber).[8]

The image files distributed by LA county were never intended to be used for data analysis. Rather, the file is simply a record of every card that was fed through the machine. Getting the data into a usable form presents a major challenge. The first challenge is mapping the computer record that represents each cast ballot to a particular pattern of holes. The next problem is matching the pattern of holes with a particular set of candidate and proposition votes. The final problem is determining which ballot images were actually used in tallying the vote (many ballots are run through the counting machine several times!). The first problem is relatively straightforward.[9] The problem of matching a pattern of hole punches with votes for particular candidates is cumbersome. In 1992, 235 different ballot layouts were used. Matching the ballot punches to votes for candidates simply requires a mapping between the punch positions and candidates for each of the 235 ballot

[7] Michael Alvarez informs me that due to a change in the computer system used to total the votes, the ballot image files are no longer being compiled.

[8] I thank Liz Gerber for making the raw data files available to me.

[9] I thank Ken McCue and Liz Gerber for providing model code for this operation.

layouts.

Determining which ballots were used in calculating the official total is the most difficult aspect of reading the data. Ballots are run through the counting machines in batches. Each batch has a header record followed by individual ballot records. The problem is further complicated because different batches are processed on several card readers simultaneously, and the log file interlaces records from the different readers. If a particular batch is not going to be used in the final tally, a card is supposed to be run through the machine that indicates that the batch will not be used. However, in practice some batches that are not counted do not have a "delete" record in the data.

Other authors have recognized this problem and have determined to treat the records they extracted from the tapes as a sample and do not reconcile the individual voting data with the precinct-level statement of the vote (Dubin & Gerber 1992 and Gerber & Many 1996). I have solved the problem of determining which vote batches were used in the final tally and am able to generate precinct-level vote totals that exactly match the official statement of the vote. First, I read through the data and extract for each batch of the votes: the precinct the batch was from, the total number of votes in the batch, and the number of valid votes for several candidates in the batch. I also note whether the batch was marked as deleted. In the 1992 election, there were over 17,000 batches used to record the votes in the 6,305 precincts. Then for each precinct, I take the reported vote total and determine which set of batches for that precinct would yield the reported vote total. If more than one set of batches match the total, the total votes for particular offices are checked until all but one set of batches has been eliminated as a potential generator of the precinct totals. Having compiled a compete (tentative) list of which ballot batches were used, I then reread the entire image file, compiling the votes for those batches that were identified as having been used in tallying the totals. As a final check, the individual-level votes are aggregated to the precinct-level and checked against the reported precinct totals.

The data presented in the paper reproduce the precinct totals to the vote for all federal and state offices and on all county and statewide propositions.

D Roll-call votes on Abortion in the US House, 1973–2000

For the 93rd through the 100th Congress, abortion votes were identified by codings given in Poole and Rosenthal's Voteview software (`http://voteview.uh.edu`). For the 101st Congress abortion votes where extracted from *Congressional Quarterly* descriptions in data provided by Jim Snyder. For the 102nd through the 106th Congress abortion related roll calls were identified through their use by the National Abortion and Reproductive Rights Action League (NARAL) and the National Right to Life Committee (NRLC) in constructing legislator ratings. The votes are identified in the Voter Information Services VIS Database (Washington DC: Voter Information Services, 2000). A comple list of all of the votes used is given in table D. For computational convenience, in years where more than 20 roll calls were taken, 20 were chosen for analysis at random.

Congress	*Congressional Quarterly vote numbers*
93	113, 183, 184, 479, 663, 664, 707, 783, 968.
94	948, 952, 1092, 1182.
95	326, 466, 550, 595, 596, 603, 675, 681, 690, 696, 701, 1087, 1088, 1290, 1344, 1436, 1496, 1515, 1516, 1521.
96	270, 288, 312, 487, 550, 629, 630, 633, 1089, 1124.
97	37, 78. 171, 711.
98	178, 334, 501, 730, 735.
99	195, 196, 197, 216, 247.
100	210, 684, 785, 835.
101	205, 277. 278, 351, 352, 568, 569, 702.
102	95, 109, 115, 147, 148, 149, 163, 222, 229, 251, 375, 380, 403, 443, 452, 458.
103	22, 66, 68, 106, 107, 157, 158, 159, 232, 306, 307, 309, 518, 534, 580, 582.
104	51, 93, 94, 125, 167, 255, 261, 307, 310, 320, 332, 349, 350, 363, 366, 382, 422, 432, 433.
105	22, 23, 61, 62, 63, 65, 75, 167, 168, 171, 217, 260, 279, 280, 284, 288, 290, 292, 321, 325.
106	173, 184, 260, 261, 301, 303, 312, 349, 350, 353, 360, 373, 464, 465, 715, 814, 929, 982, 984, 1007, 1033.

Table 3: *Roll-call votes on abortion taken in the US House, 1973–2000.*

References

Bailey, Michael. 1998. "A Random Effects Approach to Legislative Ideal Point Estimation." Presented at the 1998 Midwest Political Science Association Meeting, Chicago.

Dempster, A. P., N. M. Laird & D. B. Rubin. 1977. "Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society, Series B.* 39(1):1–38.

Lewis, Jeffrey B. 2001. "Estimating Voter Preferences Distributions from Individual-level Voting Data." *Political Analysis*. 9(3):275–297.

Londregan, John B. 2000. *Legislative Institutions and Ideology in Chile*. Cambridge University Press.

Stroud, A. H. & Don Secrest. 1966. *Gaussian Quadrature Formulas*. Englewood Cliffs NJ: Prentice-Hall.