IDENTIFICATION AND ESTIMATION OF LATENT ATTITUDES AND THEIR BEHAVIORAL IMPLICATIONS

Richard H. Spady

THE INSTITUTE FOR FISCAL STUDIES
DEPARTMENT OF ECONOMICS, UCL
cemmap working paper CWP12/06

IDENTIFICATION AND ESTIMATION OF LATENT ATTITUDES AND THEIR BEHAVIORAL IMPLICATIONS

RICHARD H. SPADY

European University Institute, Florence,
Nuffield College, Oxford, and
Centre for Microdata Methods and Practice, London

Abstract. This paper (i) formalizes conditions under which a population distribution of categorical responses to attitudinal questions ('items') has a scale representation; (ii) develops tests for whether a particular sample of item responses is consistent with a scale representation; (iii) develops methods for nonparametrically estimating the relation between an outcome and a scale value; and (iv) generalizes the foregoing to the multi-scale case. An implication of these results is that the effect of multiple latent attitudes on behaviour can be identified, even though the attitudes of an individual can never be precisely observed. We illustrate our methods using survey data from the 1992 U.S. Presidential election, where the 'outcome' is an individual's vote and the 'items' are expressions of social and policy preferences.

1. Introduction.

The notion that individuals' social and political attitudes cohere to form a consistent and stable system of belief that guides rational action is difficult to formalize in theory and confirm in practice. In this paper we set out to provide a general account that embodies the intuitive notion of 'consistent underlying attitudes' and outline methods for assessing whether a small number of such attitudes can be used to explain voting behavior, partisan identification, and similar expressions of social and political preference. The key features of our account include:

- Individuals are 'conservative' or 'liberal' on a small number of substantive scales; individuals are (weakly) ordered along these scales and there is no loss in generality in representing the population as being uniformly distributed on [0,1] on each scale or dimension.
- Among the data are elicitations of degrees of agreement with statements concerning social and political issues; the questions and the corresponding responses are called

Date: This version June 22, 2006; first version November 2, 2004.

I am grateful to Andrew Chesher, Roger Koenker, and Byron Shafer for many helpful discussions.

‘items’. (A fundamental question is whether the responses to a particular collection of items are determined by individuals’ positions on a single scale.)

- If a valid scale exists for a particular collection of items, more ‘liberal’ people (by convention, those closer to 1 on the scale) tend on average to give more liberal answers to the items defining the relevant scale. (This is what it means to be more liberal or more conservative.) A *monotonic scale representation* assigns to each point in [0,1] a probability distribution over each item’s responses that preserves the ‘liberal people on average give liberal answers’ property.
- A set of items is ‘scalable’ if there exists a monotonic scale representation that exactly reproduces the population’s joint distribution of the constituent items’ responses. Otherwise the items are not scalable.
- When, as in most practical situations, we cannot know the population item response distribution, we can use a sample of item responses to test the hypothesis that a collection of items do ‘scale’.
- An individual’s position on the [0,1] scale can never be precisely determined, not least because we have a limited number of items, each of which consists of a categorical response. Nonetheless, the conditional probability distribution of an item response with respect to a scale value is identified for the population and can be estimated from a sample. This is true whether or not the item was used in the definition of the scale.
- The preceding point can be generalized to the case where there are multiple [0,1] scales, each based on its own scalable set of items with no item appearing in more than one scale. In this case, an item response distribution that is conditional on multiple scales can be identified and estimated.

Needless to say, we are interested in expositing the above points in a manner that is free of specific parametric assumptions. To some extent, this distinguishes our approach from the psychometric literature, which is largely parametric (Skrondal and Rabe-Hesketh 2004); even nonparametric item response theory usually starts from a reduction of the item scores to their sum (Sijtsma and Molenaar 2002).

The plan of the rest of this paper is as follows. In the next section we define ‘monotonic scale representations’, the basic building block that characterizes scalable item responses, and estimate several such representations for a sample of items relating to ‘cultural’ items. In Section 3 we develop a test for whether a particular sample of item responses is compatible with a monotonic scale representation, in which case the items are said to ‘scale.’ Section 4 examines the implications of a monotonic scale representation and the limitations on inference concerning the scale position of any particular individual. Section 5 develops methods for estimating the conditional probability of an event such as 'voting for Clinton' conditional on scale value; Section 6 tackles the same problem in the presence of additional conditioning variables, such as membership in a social group. In Section 7 we briefly present results for a second scaling, based on items relating to the role of the state in assuring economic welfare, in order to provide an example for Section 8, which develops methods for the multi-scale case. Section 9 discusses some open problems, and an Appendix gives details on the estimation methods used in the main body of the paper.

2. Monotonic Scale Representation: Definition and Examples.

To give some intuition about the methods to be developed, we introduce some data drawn from a telephone survey in 1994 conducted by the Pew Foundation that explored attitudes and voting behavior in the 1992 U.S. Presidential election. We will limit our attention at this point to five items and the Presidential vote, to explore whether the five items are 'scalable,' and, if so, how the underlying attitude they manifest affects the vote.

The five items come in two formats. In the first, the respondents are read two statements and asked which of the statements best expresses their own view; after making the choice the respondent is asked: "Do you feel strongly about that, or not?". In the second format, the respondent is read a one sentence description of a policy proposal and asked whether they strongly favor, favor, oppose, or strongly oppose the proposal. Thus for each of the five items there are four possible outcomes, or 1,024 ($= 4^5$) possible combinations of responses.

The five items are:

(1) "The best way to ensure peace is through military strength." OR "Good diplomacy is the best way to ensure peace." (Format 1.)
(2) "Allowing government Medicaid benefits to help pay for abortions for low–income women." (Format 2.)
(3) "A constitutional amendment to permit prayer in public schools."
(4) "Homosexuality is a way of life that should be accepted by society." OR "Homosexuality is a way of life that should be discouraged by society."
(5) "Books that contain dangerous ideas should be banned from public school libraries." OR "Public school libraries should be allowed to carry any books they want."

The responses to these questions are coded 1 to 4, with 4 being the most liberal response. For the record, the liberal direction is to favor diplomacy, favor benefits for abortions, oppose school prayer, accept homosexuality, and oppose banning books with dangerous ideas.

We have a sample of 3,218 complete responses to the five items and to other questions such as recollection of voting behavior[1]. The sample provides sample means for each of the 1,024 cells defined by the five item responses, so for example there are 97 respondents answering {1,1,1,1,1}, the most conservative response to each item, corresponding to a (sample) mean of .03014; there are 352 empty cells. Ideally, we would like to construct a model which (1) conforms to our intuition of a (single) scale and (2) reproduces a given set of cell means or probabilities. Obviously, whether (2) can be accomplished depends on the cell probabilities we are targeting.

To address (1), we can, without loss of generality, assume that the population has an attribute u that is distributed uniformly on $[0, 1]$, with the characteristic that $u_2 > u_1$ implies the responses given by the sub-population u_2 are (weakly) more liberal (or have more 'u-ness') than those of sub-population u_1. The appropriate notion of monotonicity is (first-order) stochastic dominance: scale position u_2 is more liberal than u_1 provided the distribution of the u_2 item responses stochastically dominates the u_1 responses.

Definition 1. *Let $F_i(j; u)$ be the item i distribution function for response j at scale value u. Then the u_2 sub-population stochastically dominates the u_1 sub-population with respect to item i if $F_i(j; u_2) \leq F_i(j; u_1)$ for all j.*

Definition 2. *Item i is monotonic in scale u provided $u_2 > u_1$ implies that the u_2 sub-population stochastically dominates the u_1 sub-population with respect to item i.*

Definition 3. *Let $I = \{i_1, ..., i_k\}$ be a collection of items and $\mathcal{F}_I(\cdot; u) = \{F_{i_1}(\cdot; u), ..., F_{i_k}(\cdot; u)\}$ the corresponding collection of distribution functions. Then $\mathcal{F}_I(\cdot; u)$ is a monotonic scale representation provided (i) all elements of I are monotonic in scale u, and (ii) conditional on u, the item responses are independent so that the joint distribution function of the outcomes $F_I(j_{i_1}, j_{i_2}, ..., j_{i_k}; u) = F_{i_1}(j_{i_1}; u) \cdot F_{i_2}(j_{i_2}; u) ... \cdot F_{i_k}(j_{i_k}; u)$.*

Definition 4. *Let $I = \{i_1, ..., i_k\}$ be a collection of items admitting q distinct outcomes or responses with population probabilities $\pi = \{\pi_1, ..., \pi_q\}$. Then $\{I, \pi\}$ is a scalable item collection provided there exists a monotonic scale representation that reproduces π (i.e. assigns the probabilities π to the q possible outcomes.)*

We reiterate in passing that whether a collection of items is scalable is a property of the population probabilities π.

[1]In the current paper we consider only complete responses. Missing data and incomplete responses can be handled by fairly simple modifications of the methods developed here, but their exposition is considerably simplified once the complete response case is understood.

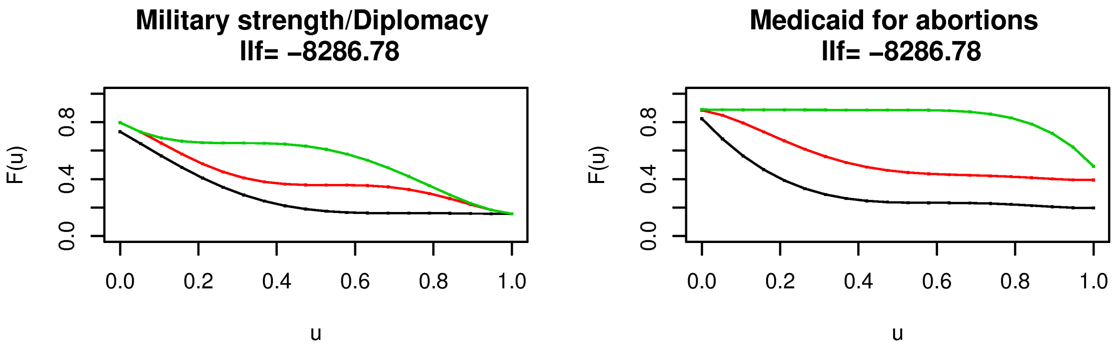

Figure 1: A monotonic scale representation for the multinomial distribution of Table 1.

To demonstrate the implications of our definitions, we consider the simple example of the collection of items 1 and 2 above. Since both items have 4 possible responses, there are 16 possible outcomes for the collection. In our sample of 3,218 responses the sample means for these outcomes are given in the following table:

Abortion→	1	2	3	4	Total
Military↓					
1	.1125	.0597	.0671	.0385	.2778
2	.0360	.0267	.0435	.0165	.1227
3	.0364	.0308	.0491	.0158	.1321
4	.1302	.0991	.1510	.0870	.4674
Total	.3151	.2163	.3108	.1579	

Table 1. Relative frequencies for items 1 and 2.

The multinomial distribution represented in Table 1 has a monotonic scale representation that is presented in Figure 1. The bottom line in each panel corresponds to the distribution function for the response “1”, the middle line to the response "2", and the top line to the response "3". Since all responses must be 4 or less, the distribution function for "4" is a horizontal line at 1. The probability of giving the response "1" is the value of the bottom line; the probability of "2" is the difference between the middle and lower lines; of "3" the difference between the top and middle lines, etc.

Several important aspects of monotonic scale representations are exemplified in Figure 1. First, monotonicity, or the property that ‘liberal people on average give liberal responses’ corresponds to the requirement that all the lines shown are (weakly) monotonically decreasing. Second, the lines may not cross, (nor go beyond [0,1]) since this would correspond to a negative probability for some value of u. Third, while the probabilities of the responses "1" and "4" must be monotonic in u, those of "2" and "3" (and in general, of any 'interior' response) need not be, as is apparent from the left panel.

Since Table 1 gives the sample frequencies of the 16 possible outcomes, its monotonic scale representation has the same value of its likelihood function as the multinomial model that assigns to the 16 outcomes their sample frequencies. While the details of the construction of Figure 1 are left to the Appendix on estimation, it is worth noting that the multinomial model provides a likelihood bound (the 'nonparametric likelihood bound') and any estimate that achieves this bound and that satisfies the monotonicity and non-negative probability constraints is also a monotonic scale representation. Consequently a monotonic scale representation need not be unique. In fact, multiple representations are possible for Table 1. However, it is not the case that every two item collection has a monotonic scale representation: reversing the coding of an item in Table 1 (i.e. recoding responses {4,3,2,1} to {1,2,3,4}) produces a multinomial distribution that cannot be monotonically scaled.

While the nonuniqueness of monotonic scale representation requires further investigation, the more important issue in practice is that the population probabilities π, which would serve as our 'target' for the estimation process, are not known but usually estimated by sample means. Since the number of empty cells in a sample grows rapidly with the number of items, (there are no empty cells in our sample of 3,218 for item collections with 16 or 64 cells, but typically 14 when there are 2 56 cells and 352 empty cells for the 1,024 possibilities when all five items are taken together), and a monotonic scale representation is (virtually) incapable of producing a probability of zero for any outcome, it will be quite common for there to be no monotonic scale representation for the sample probabilities $\hat{\pi}$ when q is large relative to n, the number of respondents. Obviously when $q > n$ (so for 6 items in the current example) this *must* happen. In addition, the growth of q relative to n increases the sampling error in $\hat{\pi}$ across an increasingly large number of its elements. Consequently, the failure of $\hat{\pi}$ to have a monotonic scale representation is not in itself evidence that the item collection $\{I, \pi\}$ is not scalable. In the next section we develop a simple test for scalability.

3. Item Selection and Tests for Scalability

A simple test for scalability follows from the observation that the sample nonparametric log likelihood value $\sum_{j=1}^{q} \hat{\pi}_j \log \hat{\pi}_j$ is an upper bound on the achievable log likelihood for a given sample and that this bound has a distribution for the data generating process represented by an estimated monotonic scale representation (or any model, for that matter.)

A test can thus be based on whether the observed value of the nonparametric log likelihood function ('npllf') is unusual for the estimated model, since an npllf that would *not* often be generated by the model is evidence that the model is overly restrictive.

This test can be implemented by simulating the estimated model for the sample size in hand and comparing the simulated values of the npllf to the value in the sample. A test of size α is constructed by referring the sample npllf to the (1-α) quantile of the simulated values, with the hypothesis of scalability being rejected if the sample npllf is more than 100f(1-α)% of the simulated values.[2]

Applying this procedure to the five items above, we obtain a npllf for the sample of -18962.98, whereas a b-spline model with df=9 (see the estimation appendix for details) has a llf of -19610.2. The estimated model has 5*3*9=135 parameters, since there are 5 items each with 3 curves to be estimated, each of which is given by a b-spline with 9 parameters. Simulating 1,000 samples from the estimated model results in simulated npllf's which exceed -18962.98 only 4.6% of the time. It is possible to increase this proportion by expanding the complexity or flexibility of the model, but only slightly. Consequently we conclude that it is unlikely that the five items scale.

Perhaps the most natural strategy to deal with this result is to reduce the number of items to be scaled to four by estimating the five models that result as each of the items is deleted in turn. The results of this process are given in Table 2:

Items		npllf	model llf	test level
1,2,3,4		-15769.79	-15936.00	.235
1,2,3,5		-15674.51	-15853.29	.255
1,2,4,5		-15708.05	-15848.79	.498
1,3,4,5		-15482.37	-15634.59	.378
2,3,4,5		-15552.14	-15729.25	.182

Table 2. Tests for scaling within subsets of four items; b-spline models with 9 parameters.

Examining this table we see there is no combination for which there is a decisive rejection of scalability, but that one combination, {1,2,4,5}, which excludes the school prayer item, seems most plausible. The estimated monotonic scale representation for this case is given in Figure 2.

An alternative to this is to alter the coding of the items. For example, it is possible to code the responses as simply 'agree' or 'disagree'; this is equivalent to removing the upper and lower lines in Figure 2, leaving only the middle line. It is possible that the resulting collection of item binary distributions is scalable while the full coding is not. Thus, to use

[2] This procedure differs from a bootstrap.

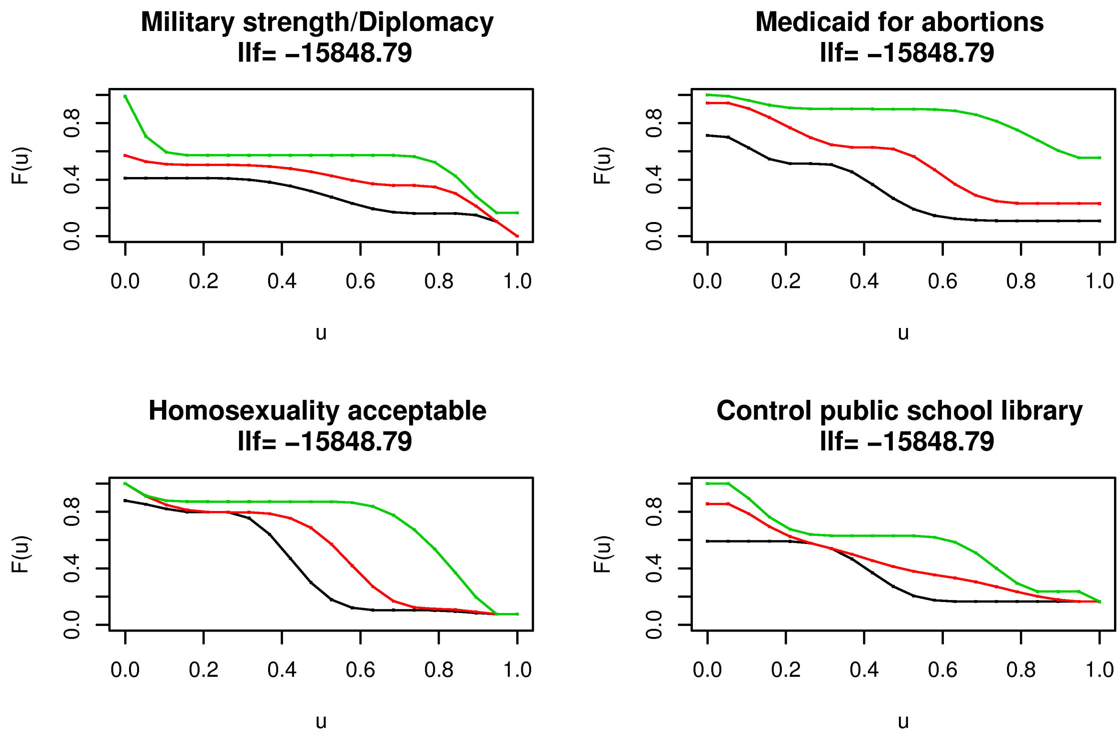

Figure 2: Monotonic scale representation for items 1,2,4, and 5. Responses are fully coded and the model has 9 b-spline terms for every line.

the binary coding is to relax the restriction of stochastic dominance in the full coding. A disadvantage of this is that individual specific information is lost. However, since the number of possible outcomes is reduced (in this example, from $4^5 = 1,024$ possibilities to $2^5 = 32$ possibilities), the sampling error in the estimated cell probabilities is reduced.

Table 3 shows the likelihoods and simulated test levels associated with a binary outcome coding for our five items for different levels of model parameterization.

b-spline terms		llf	scalability test level
4		-10189.74	.380
5		-10186.10	.405
6		-10184.79	
7		-10183.09	.456
8		-10182.82	
9		-10182.04	.470

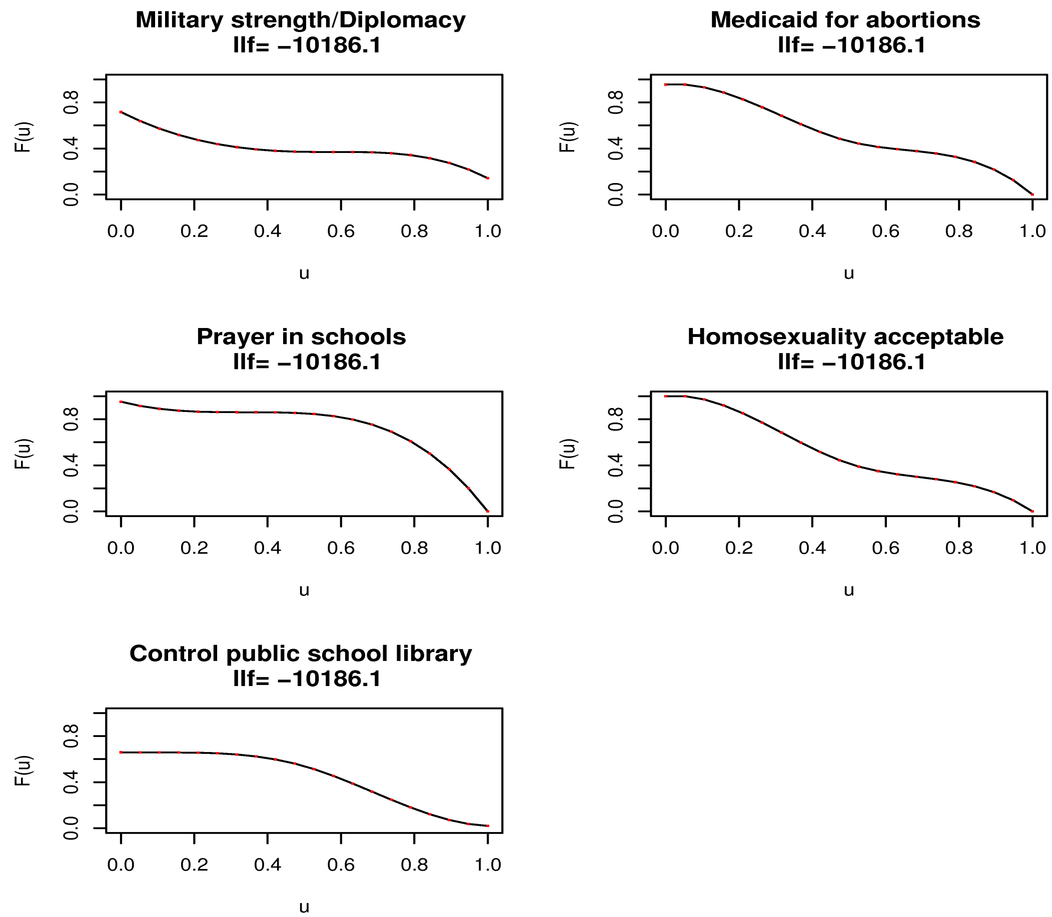

Figure 3: Monotonic scale representation for items 1,2,3,4, and 5. Responses are binary coded and the model has 5 b-spline terms.

Table 3. Likelihood values and scalability tests for binary codings of all five items.

From Table 3 it is apparent that the fit and scalability tests levels are all similar, and that the five items, when coded into two responses only, are probably scalable. We choose the model with 5 b-spline terms, illustrated in Figure 3, to subject to further analysis along with the fully coded 9 term model of Figure 2, as a potentially interesting point of contrast both with respect to coding and to the evident degree of smoothness.

4. Implications of the Estimated Monotonic Scale Representation.

The estimated monotonic scale representations provide estimates of various important quantities, such as the u–position of respondents (via the posterior distribution given their

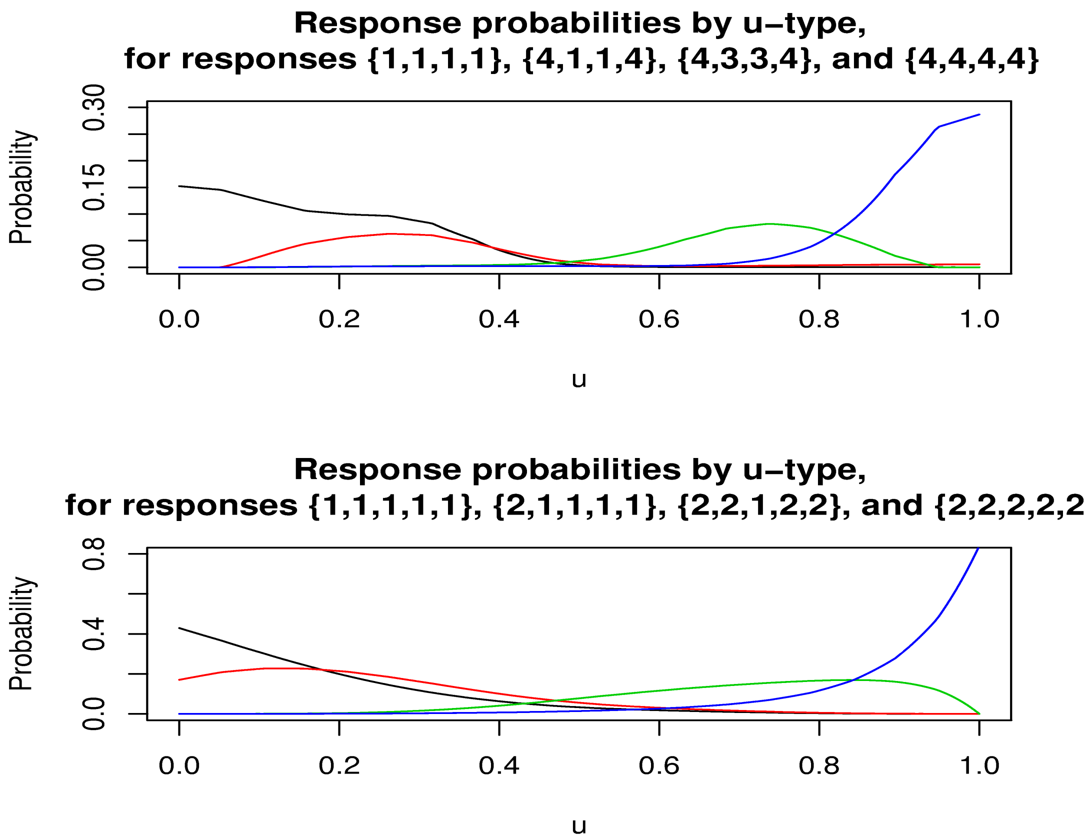

Figure 4: Response probabilities by u-type. Top panel: estimates from the fully coded four item model of Figure 2. Bottom panel: estimates from the binary coded five item model of Figure 3.

responses) and the probability of a response given a position on the u–scale. Figure 4 shows the probabilities of some common responses for the models of Figures 2 and 3. In each panel the responses that are displayed can be unambiguously ordered from conservative to liberal, and the corresponding probabilities have modes in the corresponding order, with the distributions of the extreme responses being more sharply defined (peaked) than those of the moderate responses.

How well can we infer the u–type of a respondent? To gain some sense of this, we tabulate in Table 4 the probabilities for the 32 possible responses of the binary coded model for $u = .10, .50$, and $.90$. From Table 4 it is evident that while $u = .1$ respondents quite commonly answer {1,1,1,1,1} and $u = .9$ respondents commonly answer {2,2,2,2,2}, the median (in u) respondent is spread out over a number of responses. The median respondent's third (and nearly second) most likely answer is {2,2,1,2,2}, which occurs with probability .0775, is also the second most likely answer of the $u = .9$ respondent also, who gives it with probability .1583. Consequently it is clear that we cannot, at least in this model, easily distinguish between various types of respondents on the basis of their response. When we carry out classification of respondents via Bayes' Theorem, we inevitably will attribute to some e.g. median respondents a high likelihood that they are liberals, because they have given answers that are often given by liberals.

response	$u = .1$	$u = .5$	$u = .9$	response	$u = .1$	$u = .5$	$u = .9$
11111	0.3085	0.0327	0.0002	21111	0.2247	0.0554	0.0006
11112	0.1609	0.0282	0.0029	21112	0.1172	0.0477	0.0080
11121	0.0081	0.0458	0.0011	21121	0.0059	0.0777	0.0031
11122	0.0042	0.0395	0.0154	21122	0.0031	0.0669	0.0418
11211	0.0371	0.0058	0.0004	21211	0.0270	0.0098	0.0011
11212	0.0193	0.0050	0.0055	21212	0.0141	0.0084	0.0148
11221	0.0010	0.0081	0.0021	21221	0.0007	0.0137	0.0057
11222	0.0005	0.0070	0.0286	21222	0.0004	0.0118	0.0775
12111	0.0222	0.0379	0.0008	22111	0.0162	0.0642	0.0022
12112	0.0116	0.0326	0.0112	22112	0.0084	0.0553	0.0303
12121	0.0006	0.0531	0.0043	22121	0.0004	0.0900	0.0117
12122	0.0003	0.0457	0.0584	22122	0.0002	0.0775	0.1583
12211	0.0027	0.0067	0.0015	22211	0.0019	0.0113	0.0041
12212	0.0014	0.0058	0.0207	22212	0.0010	0.0098	0.0562
12221	0.0001	0.0094	0.0080	22221	0.0001	0.0159	0.0216
12222	0.0000	0.0081	0.1083	22222	0.0000	0.0137	0.2937

Table 1: Response probabilities for $u = .1, .5, .9$ in the binary coded five item model.

Interestingly enough, though, we can use the estimated monotonic scale representation to precisely calculate this effect. For example, for the median respondent of Table 4, we know there is a .0327 probability that he will give a {1,1,1,1,1} response, a .0282 probability of a {1,1,1,1,2} response, etc. and that for each of these responses, we also know that we will compute posterior probabilities for u based on the curves analogous to those shown in Figure 4 (after normalizing by dividing by their integrals, i.e. the probability of each outcome.)

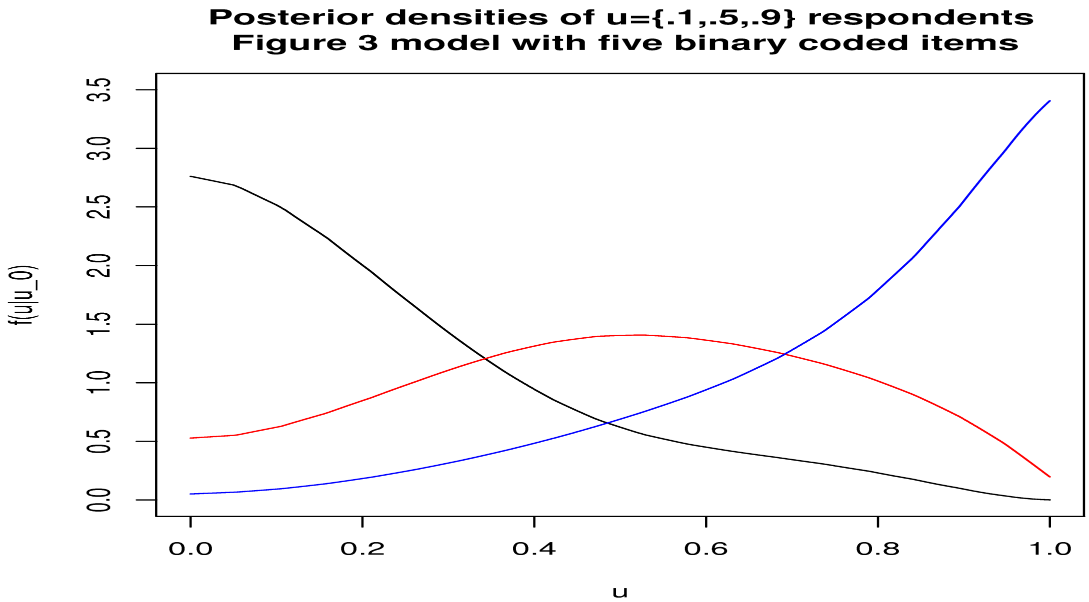

Figure 5: Posterior densities that would be attributed to populations consisting solely of $u = .1, .5, .9$ respondents, respectively. Based on the binary coded five item model.

Consequently for the median respondent we compute:

$$f(u|u_0 = .5) = \sum_{j=1}^{q} f(u|r_j)\pi(r_j|u_0 = .5) \tag{1}$$

In Figure 5 we carry out the calculation indicated in equation (1) for the median respondent and the analogous calculations for $u = .1$ and $u = .9$ respondents. One way of interpreting Figure 5 is the following thought experiment. A pure population of $u = .5$ respondents behaves as estimated by the model; we observe their responses. Using the model but not knowing anything about the input population, how would one classify the population? The limiting distribution is the middle curve in Figure 5.

One important observation from Figure 5 is that the more extreme respondents must of necessity be on average attributed toward the middle of the distribution, the extreme case being respondents at $u = 0$ or $u = 1$. This effect is operative at the $u = .1$ and $u = .9$ levels shown in Figure 5: the expectation of the corresponding posterior densities is .257 and .749, respectively; for the median it is .505.

For comparison we show the same calculations for the Figure 2 model with four fully coded items in Figure 6. Here the posteriors have modes at about the 'right' levels, but the bias in expectation is about the same: the three expectations are .259, .496, and .756,

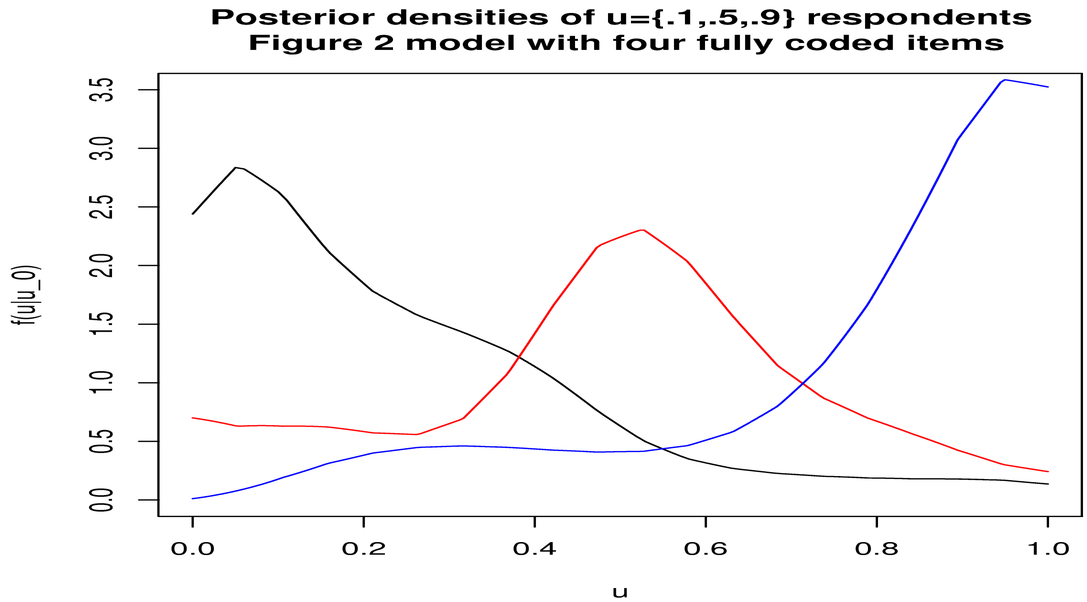

Figure 6: Posterior densities that would be attributed to populations consisting solely of $u = .1, .5, .9$ respondents for the fully coded four item model. While the modal attribution is more sharply expressed and closer to the subpopulation value, considerable 'misattribution' exists.

respectively. It could be argued that the fully coded four item model is more informative about respondents than the binary five item model, but as we shall see in the next section, the most salient features are that while classifications based on a small number of coarse items must of necessity be error–prone at the respondent level, we do know the classification structure implied by the model. This means that the 'measurement error' inflicted on us by nature can be taken into account so that relevant relations may be reliably computed at the population level. We will examine the implications of this for model selection below.

5. Estimating the Probability of an Event Conditional on a Scale Value.

We now turn attention to the simplest case of the important question of how to estimate the probability of an arbitrary event or outcome, such as a respondent's 'voting for Bush in 1992', conditional on the value of a respondent's scale value. Letting V take the value 1 if the event in question takes place and 0 otherwise, we consider this problem under two assumptions:

$$p(V|u,r) = p(V|u) \tag{2a}$$
$$p(V|u,r) = p(V|r) \tag{2b}$$

The first of these assumptions is perhaps the most natural: it says that if we knew u, knowing r would be of no additional value; in effect, the underlying characteristic u carries all the relevant information for the event V.

However, a case could be made that the second assumption is reasonable. The underlying characteristic u determines the responses r that are expressions of social and political preferences; once these have 'hardened', a decision such as the vote is framed in terms of the preferences on these items. An advantage of this approach is that it is closer to the observable data; it also, as we shall see, offers an informative benchmark.

Under the first assumption ('Model 1') we have:

$$\begin{aligned} p(V|r_j) &= \int p(V|u,r_j)f(u|r_j)du \\ &= \int p(V|u)f(u|r_j)du \end{aligned} \tag{3}$$

where $j = 1, ..., q$ indexes the q possible responses. This suggests estimating $p(V|u)$ by maximizing the conditional likelihood function across respondents.

Under the second assumption ('Model 2') the approach taken in equation (3) generates an identity since

$$\begin{aligned} p(V|r_j) &= \int p(V|u,r_j)f(u|r_j)du \\ &= \int p(V|r_j)f(u|r_j)du \\ &= p(V|r_j)\int f(u|r_j)du \end{aligned} \tag{4}$$

Instead, integrating out (summing over) the responses is called for:

$$p(V|u) = \sum_{j=1}^{q} p(V|r_j,u)p(r_j|u) \tag{5a}$$
$$= \sum_{j=1}^{q} p(V|r_j)f(u|r_j)\pi(r_j) \tag{5b}$$

Examining (5b)[3] we see that $p(V|u)$ can be estimated by using the sample analogs of $p(V|r_j)$ and $\pi(r_j)$[4] , since the item response model supplies an estimate of $f(u|r_j)$. Carrying out the indicated substitutions of sample analogs we have

$$\hat{h}(u) = \frac{\sum_{i=1}^{n} V_i f(u|r_i)}{n} \tag{6}$$

where $\hat{h}(u)$ is interpretable as an estimate of $p(V|u)$ in Model 2 and is also computable under Model 1.

Equation (6) has a further interpretation. Suppose we *knew* u_i; then the expression

$$\hat{p}(V|u) = \frac{\sum_{i=1}^{n} V_i \phi(u; u_i, \sigma)}{n} \tag{7}$$

where $\phi(u; u_i, \sigma)$ is a symmetric density function with location u_i and scale σ (or other kernel centered at u_i with bandwidth σ) would serve as a nonparametric estimate of the probability of the event V conditional on u.[5] In the current context we do not observe u_i but we have $f(u|r_i)$, its posterior distribution based on i's items responses r_i. If we substitute $f(u|r_i)$ for $\phi(u; u_i, \sigma)$ in (7) we obtain (6). This indicates that as $f(u|r_i)$ becomes more concentrated around u_i, $\hat{h}(u)$ approaches $p(V|u)$ in many circumstances.

These observations motivate another approach to estimating Model 1: to estimate $p(V|u)$ by expressing $h(u)$ as a function of $p(V|u)$ and then to estimate $p(V|u)$ by bringing the implied $h(u)$ into agreement with the observed $\hat{h}(u)$. Writing the population $h(u)$ as

$$h(u) = \sum_{j=1}^{q} p(V|r_j) p(r_j|u) \tag{8}$$

---

[3]Which follows from (5a) since $p(V|r_j, u) = p(V|r_j)$ and $f(u|r_j)\pi(r_j) = f(u, r_j) = p(r_j|u)f(u) = p(r_j|u)$.

[4]Another estimate would use the item response model's estimate of $\pi(r_j)$ rather than the sample analog.

[5]One way to see this is to see that (7) is the sample analog of

$$p(V = 1|u) = \frac{f(u|V = 1)p(V = 1)}{f(u|V = 0)p(V = 0) + f(u|V = 1)p(V = 1)}$$

we note that $p(V|r_j) = \int p(V|w)f(w|r_j)dw$, where we use w in place of u to distinguish it from the point of evaluation of $h(\cdot)$. Thus

$$\begin{aligned} h(u) &= \sum_{j=1}^{q} \int p(V|w)f(w|r_j)dw \; p(r_j|u) \qquad (9) \\ &= \sum_{j=1}^{q} \int p(V|w)f(w|r_j)p(r_j|u)dw \\ &= \int p(V|w) \sum_{j=1}^{q} f(w|r_j)p(r_j|u)dw \\ &= \int p(V|w) \sum_{j=1}^{q} f(w|r_j)f(u|r_j)\pi(r_j)dw \\ &= \int p(V|w)g(w;u)dw, \end{aligned}$$

where

$$g(w;u) = \sum_{j=1}^{q} f(w|r_j)f(u|r_j)\pi(r_j) \qquad (10)$$

The function $g(w;u)$ gives the density of 'actually–w' respondents at attribution point u; it is symmetric in its arguments. Notice that if responses were so informative that $g(u;u) = 1$ and $g(w;u) = 0$ for $w \neq u$ so that $g(w;u)$ is a Dirac delta function, then $h(u)$ would be $p(V|u)$.

An important aspect of the manipulations in (9) is that they reduce a computation with q integrals to a single integral. Moreover, if we are constructing $p(V|u)$ as part of an optimizing process, it is convenient that $g(w;u)$ depends solely on the (previously estimated) monotonic scale representation.

We now estimate $p(V|w)$ by matching $h(u) = \int p(V|w)g(w;u)dw$ with $\hat{h}(u) = n^{-1}\sum_{i=1}^{n} V_i f(u|r_i)$ for various binary events V; the details of the precise estimation methods are in the Appendix, but we basically minimize the sum of squares of $\hat{h}(u) - h(u)$ over a grid of u points.

We start by taking V to be the respondents' recollection of their votes in the 1992 elections, where the possible responses are Clinton, Bush, Perot, Did not vote. We estimate the models for the binary events {Bush, not Bush}, {Clinton, not Clinton} etc. simultaneously, thereby ensuring that the estimated probabilities of the 4 events together add to 1 and presumably also gaining statistical efficiency in the estimation. In Figure 7 we show the results obtained by approximating $p(V|u)$ by a logistic transformation of a b–spline with 4 degrees of freedom, together with the corresponding $h(u)$ and $\hat{h}(u)$. The estimates of $h(u)$ coincide closely with $\hat{h}(u)$, but this agreement can be made even closer by increasing the flexibility with which $p(V|u)$ is modeled, at the cost of rather implausible fluctuations in $p(V|u)$. This issue is discussed further in the Appendix.

From Figure 7 we can see that when $\hat{h}(u)$ is basically monotonic, the estimated $p(V|u)$ (under Model 1) is even more responsive. A simple way of looking at this is that under Model 1 the attribution process tends to move probability weight to the center of the distribution. Thus if there is an evident relation in $\hat{h}(u)$ between the scale value and the event, this relation is in fact *stronger* since the effect we are observing has been attenuated by the attribution process. If we believe the item response model, we can take account of this effect.

Examining Figure 7, we see that neither the Clinton nor the Bush $p(V|u)$ is very responsive between the .2 and .6 quantiles of u, and that the Bush $p(V|u)$ rises slightly over this range (corresponding basically to a change in participation, as evidenced in the 'Did not vote' panel.) In interpreting this result among a 'middle section' of voters, it should be kept in mind that there may be other factors—i.e. other scales—that motivate these voters (indeed all voters) and that in the multi-scale theory to be developed below, the values on these other scales need not be independent of the scale shown here (though the other scale will have a marginal $U[0, 1]$ distribution.) Thus it is entirely possible that these 'culturally moderate' voters tend to be a little more 'economically conservative' than others, and that this is the motivating factor in their voting decision.

As noted above, other approaches to estimating $p(V|u)$ are possible, particularly under Model 1. One possibility is to use

$$p(V|r_j) = \int p(V|u)f(u|r_j)du \tag{11}$$

as the basis for computing a conditional likelihood for each respondent's V, and maximizing $\sum_{i=1}^{n} \log p(V_i|r_i)$. An advantage of doing this is that ML provides a coherent framework for observation weighting (which is being done implicitly in the h–function method by the choice of u points over which the matching is done) and model selection (via likelihood ratio tests.) The results of doing this for the Bush and Clinton vote probabilities[6] are shown in Figure 8. The results are virtually identical to those obtained by the h–function method.

[6]In this case for each candidate separately, although the estimates could be done jointly as for the h function method and as for the conditional ML estimates of vote by demographic group below.

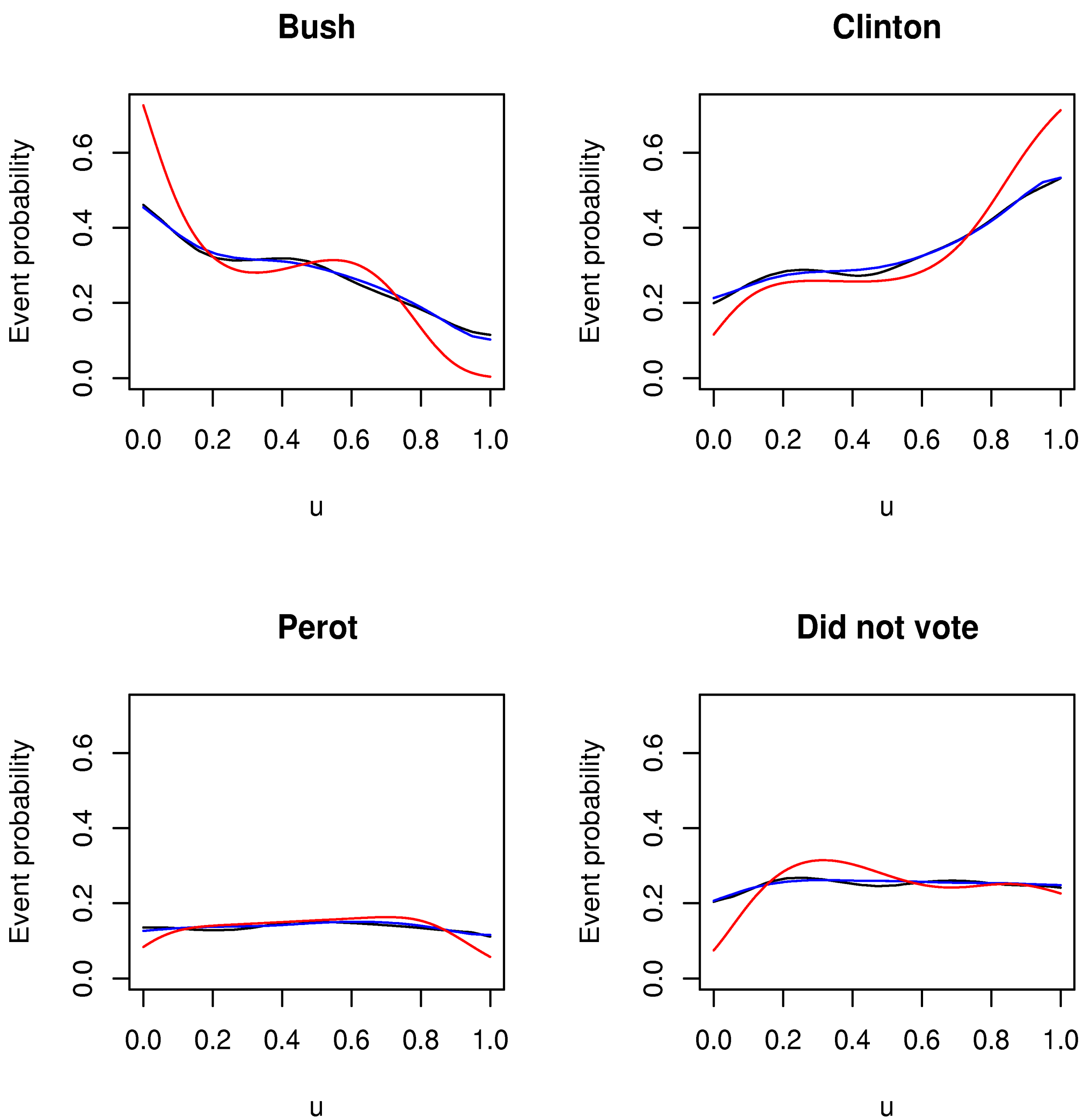

Figure 7: Probabilities of voting for Bush, etc. derived from the fully coded four item model. In each panel the two lines that are hardly distiguishable are $h(u)$ (black) and $\hat{h}(u)$ (blue). The remaining line (red) is the estimate of $p(V|u)$, which is a b–spline with four degrees of freedom.

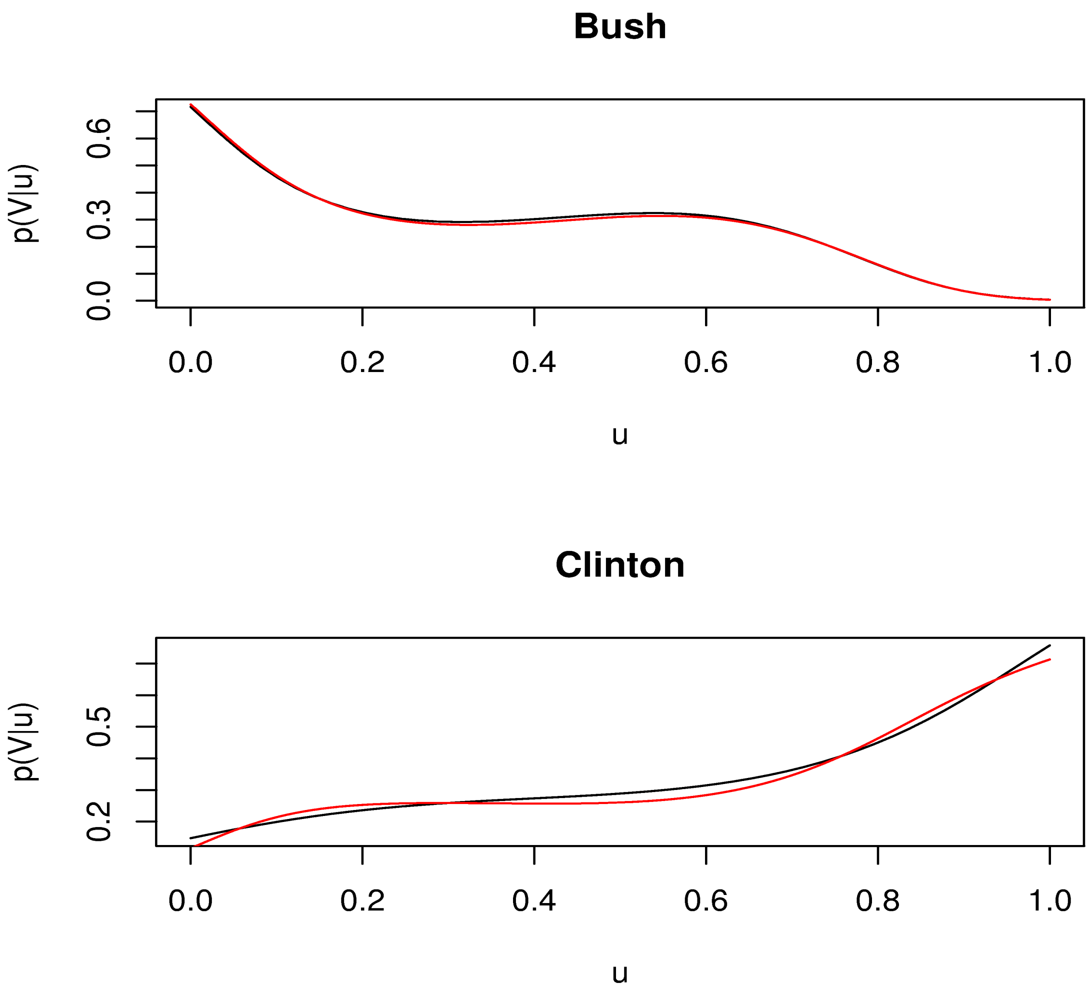

Figure 8: Conditional maximum likelihood (black line) and h function (red) estimates of the probability of voting for Bush (top panel) and Clinton (bottom).

6. Attitudes and Behavior by Social Group.

It is possible, though perhaps at first sight peculiar, to model demographic characteristics in the same way as the vote. That is, using D to denote membership in a demographic group (such as African American, college graduates, etc.) one can estimate $p(D|u)$ using the methods of the previous section. This can then serve as the basis for estimating $f(u|D)$ via

$$f(u|D) = \frac{f(u,D)}{p(D)} = \frac{p(D|u)f(u)}{p(D)} = \frac{p(D|u)}{p(D)} \tag{12}$$

since $f(u) = 1$; $f(u|D)$ gives the density of u conditional on membership in a demographic group.

We implement this idea by dividing our sample into seven mutually exclusive demographic groups following Shafer and Claggett (1996). The division is sequential, with respondents being allocated to the first group for which they qualify. The first group is African Americans; the second consists of those identifying themselves as evangelical Christians (so African Americans identifying themselves as evangelicals belong in the first group, not the second.) The third group consists of all non-Christians, including atheists. The remainder of the respondents—basically white and Asian non-evangelical Protestants and Roman Catholics—are classified by education: high school dropouts, high school graduates, those with some college or university education, and college graduates.

The proportion of our sample falling into each of the seven groups is given in Table 5 below. Five of the groups are of roughly equal size (12.2%–15.6%); Evangelicals are somewhat larger (22.7%) and Dropouts much smaller: less than 4%.

Group	Sample Proportion
African American	13.6%
Evangelical	22.7
Non-Christians	12.2
Dropouts	3.9
H.S. Grads	14.8
Some College	14.6
College Grads	15.6
Subtotal	97.3
Not Classified	2.7

Table 5. Sample proportions of demographic groups.

In Figure 9 we show the results of applying the h–function method treating demographic classification in the same fashion as V.

We can go a bit further along the same lines to extract the probability of the vote conditional on u by demographic group. Using $V * D$ to mean 'both V and D', we have

$$p(V * D|r) = \int p(V|D, u)p(D|u)f(u|r)du, \tag{13}$$

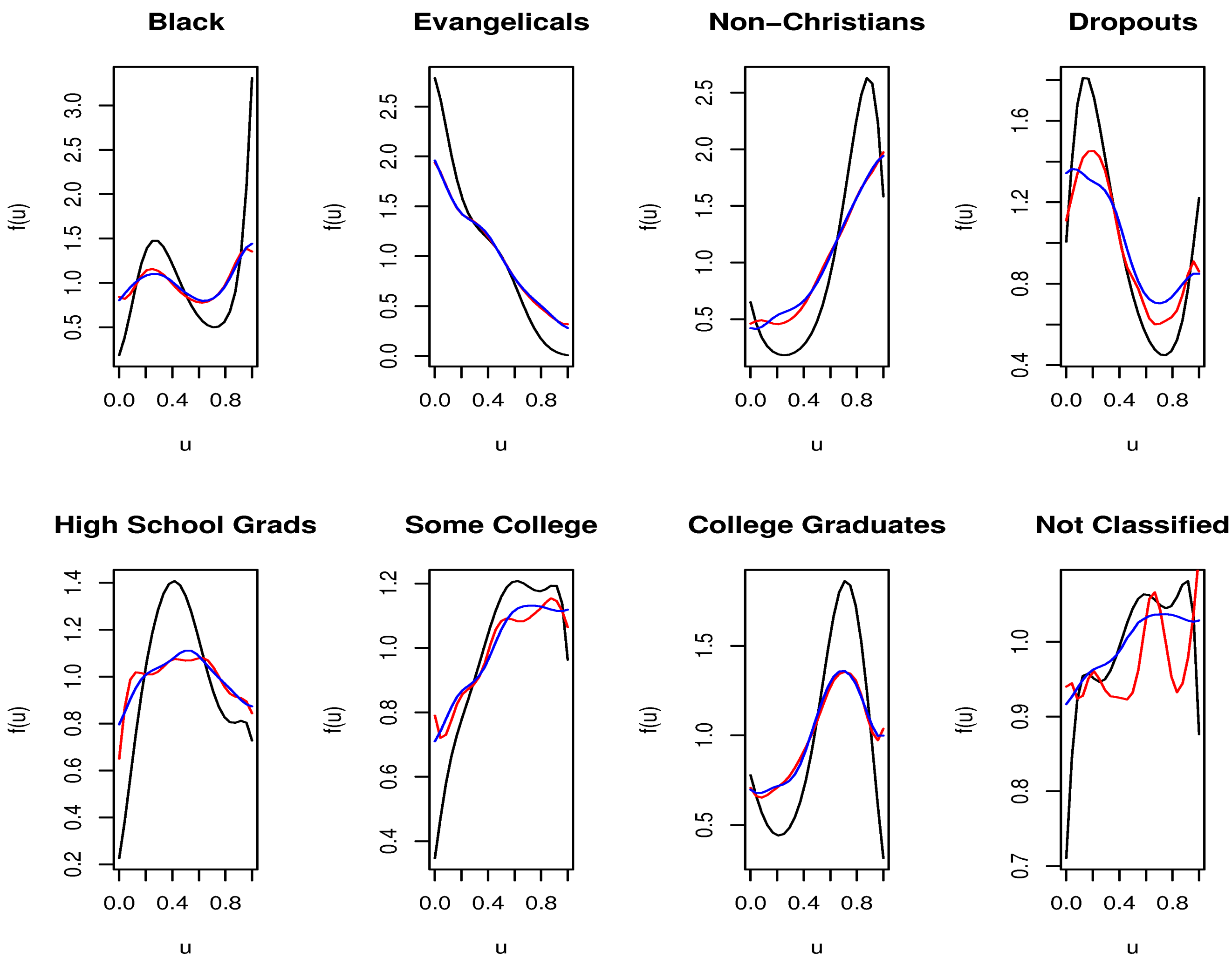

Figure 9: Estimates of $f(u|D)$ for various demographic groups based on the h–function method. The two similar lines are scaled $h(u)$ (blue) and $\hat{h}(u)$ (red).

where both $p(D|u)$ and $f(u|r)$ are previously estimated. The results of doing this for college graduates are illustrated in Figure 10.

7. A Second Scale

In order to provide an example for the multi-scale case, we develop a second scale relating to the role and efficacy of the government's assurance of economic welfare. The items from which we chose were:

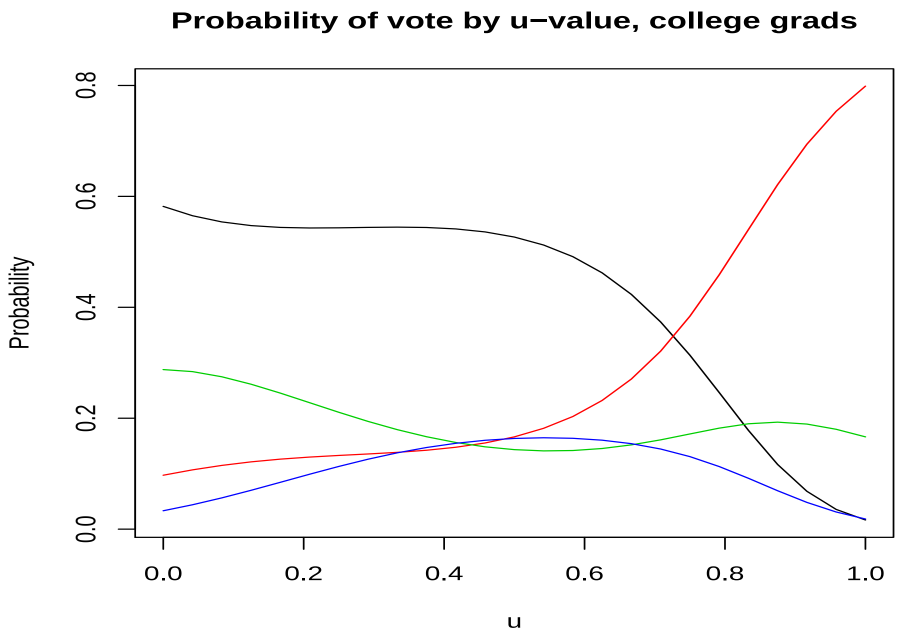

Figure 10: Probability of vote outcomes in Model 1 for 'college graduates'. The rising red line is the probability of voting for Clinton, the falling black line the probability of voting for Bush. The blue line that peacks at about $u = .5$ is the probability of not voting, while the remaining green line is the probability of voting for Perot.

(1) "Poor people today have it easy because they can get government benefits without doing anything in return." OR "Poor people have hard lives because government benefits don't go far enough to help them live decently."
(2) "Racial discrimination is the main reason why many black people can't get ahead these days." OR "Blacks who can't get ahead in this country are mostly responsible for their own condition."
(3) "Health care reform that would require employers to pay most costs of health insurance for all their workers."
(4) "New federal spending to provide education and job training for American workers whose jobs have been eliminated."
(5) "A two year limit on how long someone can receive welfare benefits."
(6) "Government is almost always wasteful and inefficient." OR "Government often does a better job than people give it credit for."

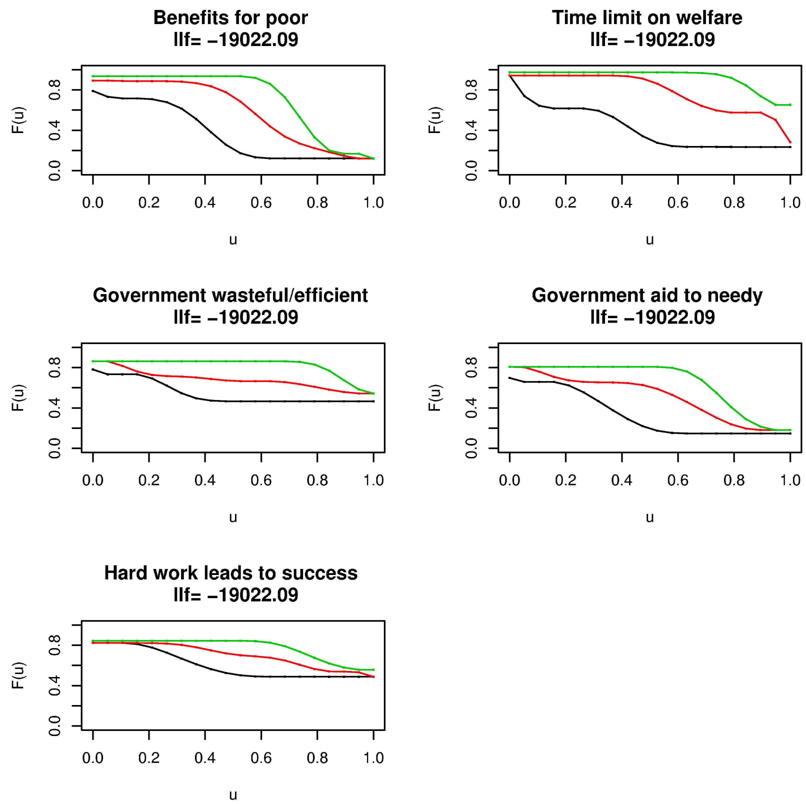

Figure 11: A monotonic scale representation for 'economic' items.

(7) "The government should do more to help needy Americans, even if it means going deeper into debt." OR " The government today can't afford to do much more to help the needy."

(8) "Most people who want to get ahead can make it if they're willing to work hard." OR "Hard work and determination are no guarantee of success for most people."

It turns out that there are a variety of combinations of items that easily scale by our criteria. For our current expository purposes, we chose the five items 1, 5, 6, 7, and 8, producing the monotonic scale representation displayed in Figure 11:

8. Multi–Scale Theory.

Suppose now that there are two scales u_1 and u_2 and that the response vector can be decomposed into two components, $r[1]$ and $r[2]$, where the items in $r[1]$ have been used to construct u_1 and those in $r[2]$ used to construct u_2, with no overlap of items. We proceed under the following assumption:

$$\begin{aligned} p(r[1]|u_1,u_2) &= p(r[1]|u_1) \\ p(r[2]|u_1,u_2) &= p(r[2]|u_2) \end{aligned} \tag{14}$$

Thus we assume that if we know u_1 then knowledge of u_2 is irrelevant to the responses pertaining to the first scale, and similarly for the second scale. Notice that this does *not* mean that $r[2]$ is uninformative about $r[1]$ or u_1, since $r[2]$ is obviously informative about u_2 and u_1 and u_2 may be 'correlated'. From (14) it follows that

$$p(r) = \int\int p(r|u_1,u_2) f(u_1,u_2) du_1 du_2 \tag{15a}$$

$$\int\int p(r[1]|u_1) p(r[2]|u_2) f(u_1,u_2) du_1 du_2 \tag{15b}$$

Since we have $p(r[1]|u_1)$ and $p(r[2]|u_2)$ from previously estimated item response models, we can estimate $f(u_1,u_2)$ by a flexible parametric or semiparametric maximum likelihood strategy, subject of course to the constraint of uniform marginals. Another possibility is to form:

$$\begin{aligned} \hat{f}(u_1,u_2) &= \sum_{i=1}^{n} \frac{f(u_1|r_i[1]) f(u_2|r_i[2])}{n} \\ &= \sum_{j=1}^{q} f(u_1|r_j[1]) f(u_2|r_j[2]) \hat{\pi}_j, \end{aligned} \tag{16}$$

where $\hat{\pi}_j$ is the sample frequency of r_j. This is a valid estimate of $f(u_1,u_2)$ under the assumptions of either Model 1 or Model 2 since $f(u_1|r_j[1])$ and $f(u_2|r_j[2])$ are consistent under the assumptions of their respective item response models. However, this estimate does not impose uniform marginals unless $\sum_{j=1}^{q} f(u_1|r_j[1])\hat{\pi}(r_j[1]) = \sum_{j=1}^{q} f(u_2|r_j[2])\hat{\pi}(r_j[2]) = 1$. While these conditions hold in both the estimated item response models and the population under the null, they need not hold in the sample. Using the values $\pi_j(r[1])$ and $\pi_j(r[2])$ as estimated by the item response models does not provide a value for $\pi(r_j[1], r_j[2])$, without making an assumption of e.g. independence. However, an adjustment can be made by dividing $\hat{f}(u_1,u_2)$ as given by (16) by $\sum_{j=1}^{q} f(u_1|r_j[1])\hat{\pi}(r_j[1]) \cdot \sum_{j=1}^{q} f(u_2|r_j[2])\hat{\pi}(r_j[2])$, which results in marginals that are nearly exactly uniform; the estimates before and after adjustment differs by no more than 1.5% and the unadjusted estimate is shown in Figure 12. The corresponding figure for the adjusted estimate is visually indistinguishable.

To extend the definition of $h(u)$ to the two scale case, consider:

$$h(u_1,u_2) = \sum_{j=1}^{q} p(V|r_j[1], r_j[2]) f(u_1|r_j[1]) f(u_2|r_j[2]) \pi(r_j[1], r_j[2]) \tag{17}$$

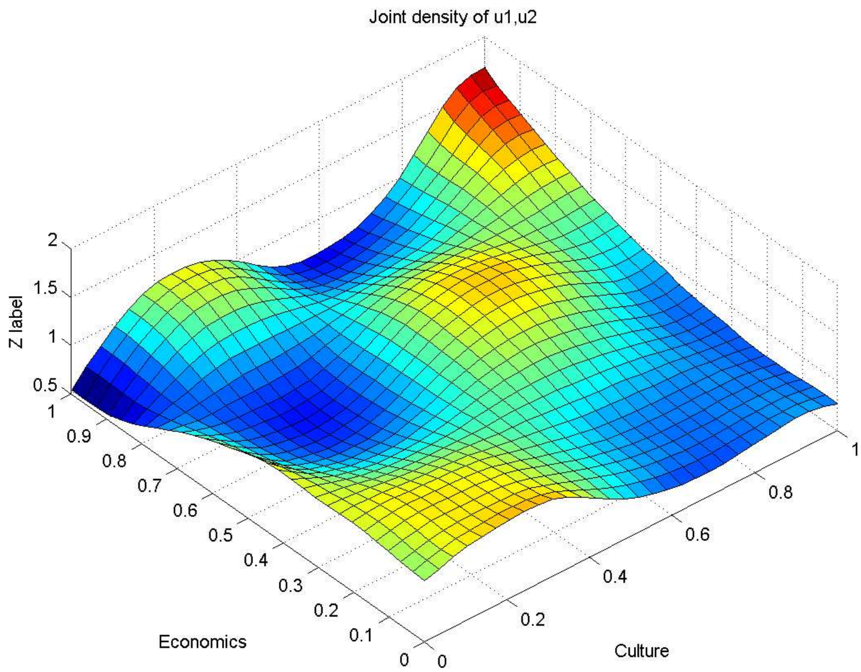

Figure 12: Estimate of the joint density of u_1 and u_2 from equation (16). Adjusting the estimate to impose uniform marginals changes the figure imperceptibly.

Since

$$f(u_1|r_j[1]) = \frac{p(r_j[1]|u_1)}{\pi(r_j[1])},$$

etc. we can follow the exposition of Section 6 by writing

$$\begin{aligned} h(u_1, u_2) &= \sum_{j=1}^{q} \int\int p(V|w_1, w_2) f(w_1, w_2|r_j) dw_1 dw_2 p(r_j[1]|u_1) p(r_j[2]|u_2) \frac{\pi(r_j)}{\pi(r_j[1])\pi(r_j[2])} \\ &= \sum_{j=1}^{q} \int\int p(V|w_1, w_2) f(w_1, w_2|r_j) p(r_j[1]|u_1) p(r_j[2]|u_2) \frac{\pi(r_j)}{\pi(r_j[1])\pi(r_j[2])} dw_1 dw_2 \\ &= \int\int p(V|w_1, w_2) \sum_{j=1}^{q} f(w_1, w_2|r_j) p(r_j[1]|u_1) p(r_j[2]|u_2) \frac{\pi(r_j)}{\pi(r_j[1])\pi(r_j[2])} dw_1 dw_2 \\ &= \int\int p(V|w_1, w_2) \sum_{j=1}^{q} f(w_1, w_2|r_j) f(u_1|r_j[1]) f(u_2|r_j[2]) \pi(r_j) dw_1 dw_2 \qquad (18) \end{aligned}$$

Defining

$$g(w;u) = \sum_{j=1}^{q} f(w_1, w_2|r_j) f(u_1|r_j[1]) f(u_2|r_j[2]) \pi(r_j) \tag{19}$$

we can write (18) compactly as

$$h(u_1, u_2) = \int\int p(V|w_1, w_2) g(w;u) dw_1 dw_2 \tag{20}$$

Recognizing that

$$\begin{aligned} f(w_1, w_2|r_j) &= \frac{p(r_j[1]|w_1, w_2) p(r_j[2]|w_1, w_2) f(w_1, w_2)}{\pi(r_j)} \\ &= \frac{p(r_j[1]|w_1) p(r_j[2]|w_2) f(w_1, w_2)}{\pi(r_j)}, \end{aligned}$$

equation (19) simplifies to

$$g(w;u) = \sum_{j=1}^{q} f(w_1|r_j[1]) f(w_2|r_j[2]) f(u_1|r_j[1]) f(u_2|r_j[2]) \pi(r_j[1]) \pi(r_j[2]) f(w_1, w_2) \tag{21}$$

which is now seen to be entirely analogous to the one dimensional case given in equation (10) after recognizing that $f(w_1, w_2)$ is $f(w)$ in the one dimensional case, which is identically 1. With the exception of $f(w_1, w_2)$, which can be estimated as outlined above, all the expressions in (21) can be derived from the estimated item response models and/or sample analogs. Moreover, this argument apparently extends to further dimensions so that equations (20) and (21) generalize to higher dimensions.

We can construct $\hat{h}(u_1, u_2)$ for the events vote Bush, vote Clinton, vote Perot, did not vote; this is done in the following figures, which represent estimates of the probability of each event under Model 2. We have not yet attempted to calculate the corresponding figures for Model 1, which must almost certainly show more dramatic effects.

9. Some Open Problems.

Our discussion has not included a rigorous treatment of identification but has instead focused on formal derivations of probabilities, densities, and likelihoods. In our computations we have employed a liberal dose of regularization by employing splines and sometimes linear interpolations between spline determined points. Some experimentation has convinced us that the estimated monotonic scale representations presented here are relatively insensitive to the degree of regularization; this is less true for the h–function method. These observations lead to at least two points that need further investigation. Both may be understood as issues in regularization.

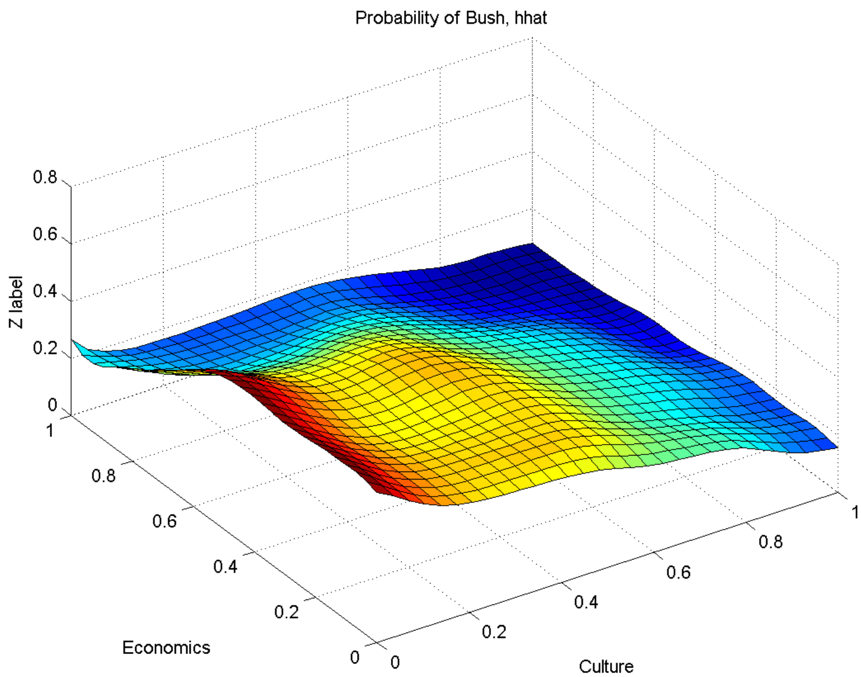

Figure 13:

(1) Particularly for small q, i.e. a small number of alternatives, there is unlikely to be a unique monotonic scale representation. For large q the problem is hidden because there will be no monotonic scale representation for the sample frequencies, and so it is natural to treat the problem as one of model selection, of devising a strategy for best choosing the effective number of parameters in a flexible model that represents the data.

(2) Even given a particular monotonic scale representation, implementations of Model 1 can be made to lean very heavily on features of the $\hat{h}$ function that are probably accidental. These implementations are almost certainly overly parameterized. It may be possible to give likelihood based procedures for model selection here since conditional maximum likelihood appears to always be a feasible estimation method for these problems.

References

Blundell, R., and J. Horowitz (October, 2004): "A Nonparametric Test of Exogeneity," Working paper, University College, London, and Northwestern University.

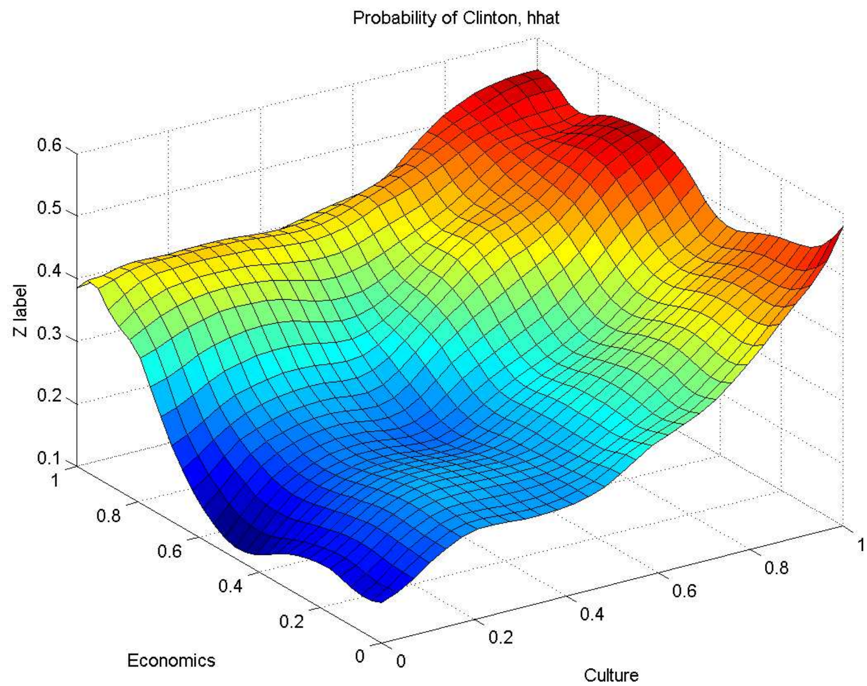

Figure 14:

Hu, Y. (August, 2004): "Identification and Estimation of Nonlinear Models with Misclassification Errors Using Instrumental Variables: A General Solution," Working paper, University of Texas at Austin.

Lewbel, A. (1998): "Semiparametric Latent Variable Model Estimation with Endogenous or Mismeasured Regressors," *Econometrica*, 66, 105–121.

Shafer, B. E., and W. J. Claggett (1995): *The Two Majorities: The Issue Constext of Modern American Politics*. The Johns Hopkins University Press, Baltimore and London.

Sijtsma, K., and I. W. Molenaar (2002): *Introduction to Nonparametric Item Response Theory*. Sage Publications, Thousand Oaks, CA.

Skrondal, A. W., and S. Rabe-Hesketh (2004): *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Princeton University Press, Princeton, N.J.

Appendix: Estimation Methods.

Estimation Methods for Item Response Models.

The item response models are estimated by maximum likelihood, subject to the constraint that the distribution functions (the lines illustrated in Figures 1, 2, 3, and 11) be downward sloping and not cross. The distribution functions are computed at 20 grid points on a b-spline and linearly interpolated between the points. The resulting probabilities for

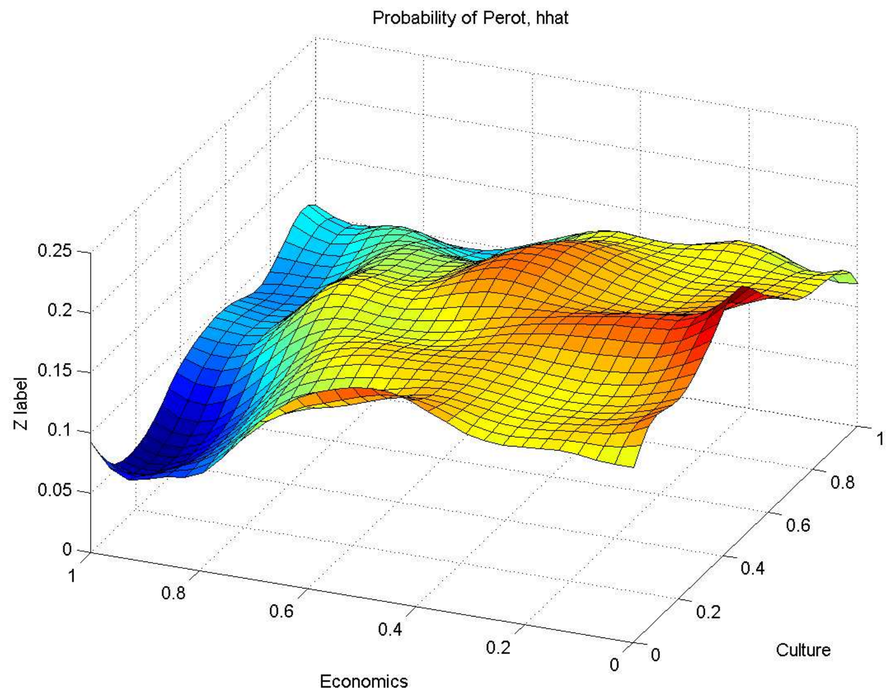

Figure 15:

a single item response are piecewise linear and the likelihood for a particular outcome or response for k items is the product of k probabilities, so the probability as a function of u is a piecewise k–degree polynomial. This is exactly integrated to evaluate the likelihood. There is no discernible difference in carrying out the computations on 50 or even 100 grid points.

European University Institute, Florence

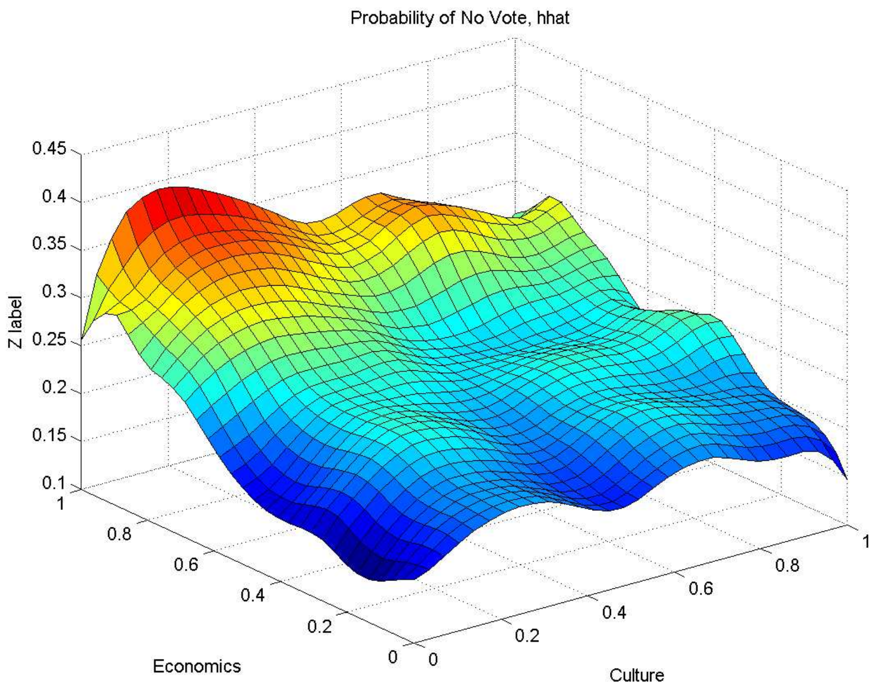
Probability of No Vote, hhat
Z label
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
1
0.8
0.6
0.4
0.2
0
Economics
0
0.2
0.4
0.6
0.8
1
Culture

Figure 16: