Journal of Data Science **6**(2008), 155-171

# Recovering Vote Choice from Partial Incomplete Data

Wendy K. Tam Cho[1] and George G. Judge[2]
[1] *University of Illinois at Urbana-Champaign*
*and* [2] *University of California, Berkeley*

*Abstract*: In voting rights cases, judges often infer unobservable individual vote choices from election data aggregated at the precinct level. That is, one must solve an ill-posed inverse problem to obtain the critical information used in these cases. The ill-posed nature of the problem means that traditional frequentist and Bayesian approaches cannot be employed without first imposing a range of assumptions. In order to mitigate the problems resulting from incorporating potentially inaccurate information in these cases, we propose the use of information theoretic methods as a basis for recovering an estimate of the unobservable individual vote choices. We illustrate the empirical non-parametric likelihood methods with some election data.

*Key words:* Cressie-Read divergence measure, ill-posed inverse problem, information theoretic methods, maximum entropy, voting behavior, voting rights act.

## 1. Introduction

Forty years ago, Congress passed the landmark Voting Rights Act (VRA) in a monumental effort to safeguard and protect the voting rights of all U.S. citizens, regardless of race or color. Indeed, soon after passage of the VRA, black voter registration increased sharply. Although the VRA has been called the single most effective piece of civil rights legislation ever passed by Congress, significant issues of enforcement remain. One particular problem occurs repeatedly after the decennial redistricting when minority advocacy groups invariably find districts they claim are in violation of the VRA because they do not allow minorities to elect the representative of their choice.

Since these are legal claims, the judge is the final arbiter. He must decide whether minority voting strength has been diluted so that minorities are not able to elect the representative of their choice. To rule in vote dilution cases, the judge must have an estimate of the unobservable relative proportions of each racial group that voted for each candidate in the district where only the total numbers in each racial group and the total votes won by each candidate are

known. Thus, the secret ballot poses an obvious difficulty, since we can never know with certainty how a particular voter or group of voters voted. Instead, we have only aggregated election returns and census data on racial composition for areal units as a basis for inferences about the individual-level behavior. In order to recover estimates of the unobservable quantities of interest, we must use indirect aggregate data. This results in an ill-posed pure inverse problem that is commonly known as the ecological inference problem (Goodman, 1953). Judges confronted with this problem cannot simply shrug off the considerable difficulties of ecological inference. They must somehow choose between competing accounts that likely feature different theories, models, and estimates. The stakes, likewise, are high, demonstrating the enormous role that model formation, estimation, and inference play in defining democracy in the United States.

The formulations are recognized as inverse problems because one must use indirect observations to recover information about the focus of interest, the unobservables. The problem is ill-posed or underdetermined because there are more unknowns than data points and thus insufficient information to solve the problem uniquely. These considerations suggest that these types of ecological inference problems that define the issues in such areas as Voting Rights cases are not amenable to frequentist and Bayesian estimation and inference procedures without the imposition of strong assumptions.

In this paper, we demonstrate an information-theoretic basis for information recovery in problems for ecological inference that has the virtue that it rests on a family of minimum distance criterion functions including empirical likelihood and maximum entropy principles. These methods are especially useful to process and recover information when the only available data are partial and incomplete. Traditional attempts to solve the ecological inference problem in a sampling theory and Bayesian context are illustrated by (Goodman, 1953, Goodman, 1959, King, 1997, Wakefield, 2004). Goodman (1953, 1959) assumes a sampling framework and converts the ill-posed problem to a well-posed inverse problem with noise. Each of these distinct models impose questionable assumptions and so share the statistical and social consequences of imposing strong and possibly incorrect assumptions. Building on the creative efforts of these and others, we focus on the informational content of the data, acknowledge the inherent uncertainty of the problem, and employ empirical non-parametric likelihood methods that minimize the use of information that a researcher does not possess.

We proceed as follows. First, using only the available aggregate election returns, we demonstrate how to formulate voting rights cases as a pure ill-posed inverse problem. Next, we show how information theoretic procedures provide a basis for recovering estimates of vote choice. We then demonstrate how our approach is especially useful in VRA cases where judges often insist on a point

estimate to guide their decision making on a micro-precinct basis. Finally, we conclude with a discussion of generalizations that are a basis of current research.

## 2. Statement of the Problem

We may formalize the problem in voting rights cases as follows. Consider the observed outcomes for a particular election across $i = 1, \ldots, m$ electoral units (e.g., precincts, parishes, districts, etc.). Each unit has $j = 1, \ldots, g$ types of individual voters and $k = 1, \ldots, c$ candidates for office (which may include an abstention category). Assume without loss of generality that the election units are precincts. For each precinct, the observed information is the number of votes each candidate received, $N_{i\cdot k} = \sum_{j=1}^{g} N_{ijk}$, and the number of voters in each group, $N_{ij\cdot} = \sum_{k=1}^{c} N_{ijk}$. The total number of ballots cast in the precinct is $N_i = \sum_{j=1}^{g} \sum_{k=1}^{c} N_{ijk}$. Because of the secret ballot, the total number of votes cast by each group for particular candidates in the election is unknown and unobserved. Given the observed data, our initial objective is to formulate a pure inverse model that will permit us to estimate $N_{ijk}$, the unobserved number of votes cast in precinct $i$ by voters of type $j$ for candidate $k$, from the aggregated election returns.

Table 1: Known and Unknown Components in Voting Rights Cases

| **Canidate** | **1** | **2** | **3** | **4** | **Count** |
|---|---|---|---|---|---|
| **Group 1** | $p_{11}N_{1\cdot}$ | $p_{12}N_{1\cdot}$ | $p_{13}N_{1\cdot}$ | $p_{14}N_{1\cdot}$ | $N_{1\cdot}$ |
| **Group 2** | $p_{21}N_{2\cdot}$ | $p_{22}N_{2\cdot}$ | $p_{23}N_{2\cdot}$ | $p_{24}N_{2\cdot}$ | $N_{2\cdot}$ |
| **Group 3** | $p_{31}N_{3\cdot}$ | $p_{32}N_{3\cdot}$ | $p_{33}N_{3\cdot}$ | $p_{34}N_{3\cdot}$ | $N_{3\cdot}$ |
| | $N_{\cdot 1}$ | $N_{\cdot 2}$ | $N_{\cdot 3}$ | $N_{\cdot 4}$ | $N$ |

Our data may be expressed in terms of the observed row or column proportions. That is, for precinct $i$, $n_{i\cdot k} = N_{i\cdot k}/N_i$ or $n_{ij\cdot} = N_{ij\cdot}/N_i$. The pure inverse problem may be equivalently stated in terms of the proportion of voters in each category, $p_{ijk} = N_{ijk}/N_{ij\cdot} = n_{ijk}/n_{ij\cdot}$, where $\sum_{k=1}^{c} p_{ijk} = 1$ for each $i$ and $j$. In this context, $p_{ijk}$ is the conditional probability that voters in precinct $i$ and group $j$ voted for candidate $k$, where the conditioning indices are $i$ and $j$. Table 1 illustrates the problem for a single precinct. Given the information in the margins of Table 1, we wish to recover the information in the cells of the Table. In the Voting Rights arena, the index $j$ represents racial groups, and attention may be directed chiefly at black versus non-black voting behavior. The objective in these cases is to estimate the conditional probability that a voter selected candidate $k$

given that he is a member of racial group $j$.[1] Given the observed data and the unknown and unobservable parameters for the problem may be summarized in the form of a contingency table, where the observable aggregate data are reflected in row and column sums and the unknown or unobservable data are the conditional probabilities in the interior cells of the table (see Good, 1963, Gokhale and Kullback, 1978).

### 2.1 Modeling voting behavior as an ill-posed pure inverse problem

Consider Table 1 where the observed number of ballots cast by registered voters in each group ($N_{j\cdot}$) are the row sums, and the observed number of votes received by each candidate ($N_{\cdot k}$) are the column sums. What we do not know and cannot observe is the number of votes cast by each group, $N_{jk}$, or the proportion of votes cast by each group for each candidate, $n_{jk}$. If the conditional probabilities, $p_{jk}$, were known, we could derive the unknown number of voters as $N_{jk} = p_{jk}N_{j\cdot}$. However, because the conditional probabilities are unobserved and not accessible by direct measurement, we are faced with an inverse problem where we must use aggregate data to recover the unknown conditional probabilities.

Some structure is provided by the realization that the conditional probabilities, $p_{jk}$, must satisfy the additivity condition, $\sum_{k=1}^{c} p_{jk} = 1$, and the column sum conditions, $\sum_{j=1}^{g} p_{jk}N_{j\cdot} = N_{\cdot k}$. The column sum conditions give us the relationship

$$n_{i\cdot k} = \sum_{j=1}^{g} n_{ij\cdot}p_{ijk} \ ,$$

for $i = 1, \ldots, m$ and $k = 1, \ldots, c$. To formalize our notation, we let $\mathbf{x}(i) = (n_{i1\cdot}\ n_{i2\cdot}\ \cdots\ n_{ig\cdot})'$ represent the $(g\times 1)$ vector of proportions for each of the voter groups $j = 1, \ldots, g$ in precinct $i$, and let $\boldsymbol{y}(i) = (n_{i\cdot 1}\ n_{i\cdot 2}\ \cdots\ n_{i\cdot c})'$ represent the $(c \times 1)$ vector of vote proportions for each candidate $k = 1, \ldots, c$, in precinct $i$. Then, the relationship among the observed marginal proportions and unknown conditional probabilities may be written as

$$\boldsymbol{y}'(i) = \mathbf{x}'(i)\mathbf{P}(i) \ . \tag{2.1}$$

The component $\mathbf{P}(i) = (\boldsymbol{p}_{i1}\ \boldsymbol{p}_{i2}\ \cdots\ \boldsymbol{p}_{ic})$ is an unknown and unobservable $(g \times c)$ matrix of conditional probabilities and $\boldsymbol{p}_{ik} = (p_{i1k}\ p_{i2k}\ \cdots\ p_{igk})'$ is the $(g\times 1)$ vector of conditional probabilities associated with precinct $i$, and candidate $k$. If we

[1] It is important to note that, for expository purposes, we have organized the data into a $3 \times 4$ table. Our interest is in the general $r \times c$ case where $r$ and $c$ are not restricted. This is a departure from much of the previous applied work in this area, which has revolved around problems that are definable by a $2 \times 2$ table. That limitation is unduly restrictive and, for most real-world situations, unrealistic.

rewrite $\mathbf{P}(i)$ in $(gc \times 1)$ vectorized form as $\boldsymbol{p}(i) = \text{vec}(\mathbf{P}(i)) = (\boldsymbol{p}'_{i1}\ \boldsymbol{p}'_{i2}\ \cdots\ \boldsymbol{p}'_{ic})'$, then we may, in the case of $m$ precincts, rewrite (2.1) as

$$\begin{bmatrix} \boldsymbol{y}(1) \\ \boldsymbol{y}(2) \\ \vdots \\ \boldsymbol{y}(m) \end{bmatrix} = \begin{bmatrix} \mathbf{x}'(1) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}'(2) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}'(m) \end{bmatrix} \begin{bmatrix} \boldsymbol{p}(1) \\ \boldsymbol{p}(2) \\ \vdots \\ \boldsymbol{p}(m) \end{bmatrix} \tag{2.2}$$

or in compact form as

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{p}. \tag{2.3}$$

We interpret (2.1) as a framework for information recovery at the precinct level and (2.2) as a framework for a set of precincts, such as a congressional district. The formulations in (2.1) and (2.2) connecting the unknown and unobservable voter proportions are in the form of pure ill-posed inverse problems. Given inverse problems (2.1) or (2.3), $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}$, where $\boldsymbol{y} = (y_1, y_2, \ldots, y_k)$ is a finite-dimensional observation vector, $\mathbf{X}$ is a known linear operator that is *non*-invertible, and $\boldsymbol{\beta}$ is an *unknown* high-dimensional parameter vector. The inverse problem is to recover the unobservable $\beta_i$s based on the observations, $\boldsymbol{y}$ and $\mathbf{X}$. This general formulation captures a frequently occurring problem where a function must be inferred from insufficient information that specifies only a feasible or plausible set of functions or solutions. In other words, this is a *pure ill-posed inverse problem* that is fundamentally underdetermined and indeterminate because there are more unknown and unobservable parameters than data points on which to base a solution. Consequently, prima facie, insufficient sample information exists to solve the problem using traditional rules of logic.

## 3. Information Theoretic Formulation and Solution

The two basic components needed to implement the model of voting behavior introduced in Section 2 are the data and a criterion or objective function. In VRA cases, we have already discussed the issues surrounding the first component, data. At this juncture, many have chosen to apply sampling theory or Bayesian criteria such as a maximum likelihood or least squares as a basis for estimation and inference. We diverge from this common course because, given the ill-posed pure inverse nature of the information recovery problem, these frameworks are not applicable without the imposition of a large number of strong assumptions. Instead, within the nature of our conditional probability recovery problem, we integrate the notion of relative entropy or Kullback-Leibler divergence (see Jumarie, 1990). Our method highlights the information theory contributions of Claude Shannon (1948, 1949). Shannon began with an entropy measure of uncertainty in a random

variable $\boldsymbol{Y}$ assuming a finite number of values, $y_1, y_2, \ldots, y_n$, with probabilities, $p_1, p_2, \ldots, p_n$. He then defined the uncertainty or information, $H(\boldsymbol{Y})$, in $\boldsymbol{Y}$ as

$$-H(\boldsymbol{Y}) = p_1 \log p_1 + \cdots + p_n \log p_n \tag{3.1}$$

Interestingly, Shannon's information theory does not deal with information per se but with data—the raw material from which information is obtained.

A far reaching generalization of Shannon's Theory is the maximum entropy principle enunciated by (Jaynes, 1957). The maximum entropy (MaxEnt) principle or criterion favors, out of all distributions consistent with a given set of constraints (data), the distribution that maximizes entropy. Together, information theory and the MaxEnt principle provide a basis for the investigation of all types of systems without the need to understand the relations underlying the probabilities. Note that the MaxEnt principle or criterion is a member of the Cressie-Read (1984, 1988) family of minimum divergence distance measures. The Cressie-Read power-divergence (CR) statistic (Cressie and Read, 1984; Read and Cressie, 1988; Baggerly, 1998)

$$I(\boldsymbol{p}, \boldsymbol{q}, \lambda) = \frac{2}{\lambda(1+\lambda)} \sum_i p_i \left[ \left( \frac{p_i}{q_i} \right)^{\lambda} - 1 \right] , \tag{3.2}$$

provides a distance or discrepancy measure between $\boldsymbol{p}$ (i.e., the conditional probabilities in our problem) and a set of reference weights $\boldsymbol{q}$. The discrete weights must satisfy $(p_i, q_i) \in (0,1) \times (0,1)\ \forall\ i$ and $\sum_i p_i = \sum_i q_i = 1$, conditional on the choice of $\lambda$. The CR distance measure (3.2) encompasses a family of objective functions that includes the Kullback-Leibler cross-entropy, $\sum p_i \ln(p_i/q_i)$ (Kullback, 1959; Gokhale and Kullback, 1978), the Shannon entropy functional, $-\sum p_i \ln(p_i)$ (Jaynes, 1957), the empirical likelihood criterion, $\sum \ln(p_i)$ (Burg, 1967; Owen, 1988; Owen, 1990), and the log Euclidean measure, $\sum_i (p_i^2 - 1)$. A natural default for these reference weights in the absence of auxiliary information, Laplaces's principle of insufficient reason, is a uniform distribution. For a more complete discussion of the CR statistic and corresponding family of criterion function, see Mittelhammer, Judge and Miller (2000).

In the formulations and analyses to follow, we make use of the Shannon/Jaynes entropy criterion, $-\sum p_i \log(p_i)$. In this case, given data in the form of (2.1) or (2.3), we seek estimates of the conditional probabilities, $p_{ij}$, that maximize (3.1) subject to (2.1) or (2.3) and the additivity and column sum conditions. Note that probabilities are, by definition, nonnegative, and this nonnegativity condition is seamlessly fulfilled in our model since the probabilities are expressed as an exponential function of the parameters and data. The maximum entropy criterion is an appealing solution for two reasons. First, it embodies a minimum number of

assumptions and is thus an instance of the principle of Occam's razor, i.e. it seeks to avoid extraneous information that might introduce inconsistencies, biases, ambiguities, or redundancies. Second, it produces a maximum multiplicity solution, which means that the resulting estimate is the set of conditional probabilities consistent with our data constraints that can be realized in the greatest number of ways. In any estimation and inference problem, it is crucial to be able to separate the role that the sample information plays from the role that the statistical model specification plays in determining the results. The ability to do this is one of the particular appeals of the MaxEnt approach. In the formulations ahead, our focus is on recovering point estimates of the unknown conditional probabilities.

### 3.1 The MaxEnt voter response formulation

Typically, the only available information in the voter response problem is the data contained in the margins of Table 1. If we make use of this aggregate information, under the Kullback-Leibler (following Shannon and Jaynes) estimation criterion, the pure inverse model (2.1)–(2.3) may be formulated as

$$\arg\min_{p_{ijk}} \sum_{i=1}^{m}\sum_{j=1}^{g}\sum_{k=1}^{c} p_{ijk}\ln\left(p_{ijk}/q_{ijk}\right) , \tag{3.3}$$

subject to the column-sum condition,

$$n_{i\cdot k} = \sum_{j=1}^{g} n_{ij\cdot}p_{ijk} \tag{3.4}$$

and the additivity condition,

$$\sum_{k=1}^{c} p_{ijk} = 1 \qquad \forall\ i, j\ . \tag{3.5}$$

In this way, the problem is simply stated as a constrained minimization problem that minimizes the distance between the estimated $p_{ijk}$ and $q_{ijk}$, a reference distribution. If we are agnostic, the conventional choice of $q_{ijk}$, and the one we employ, is the uniform distribution. Obviously, depending on the external knowledge base, other fixed or random $q_{ijk}$ may serve as the reference distribution. Note that the statement of the pure inverse problem, in an extreme context, involves three components: the goodness-of-fit measure (3.3), the data constraint (2.1) or (2.3) in the form of (3.4), and the additivity condition (3.5).

The Lagrangian function for the constrained minimization problem expressed in (3.3)–(3.5) is

$$L(\boldsymbol{p},\boldsymbol{q},\boldsymbol{\alpha},\boldsymbol{\gamma}) = \sum_{i=1}^{m}\sum_{j=1}^{g}\sum_{k=1}^{c} p_{ijk}\ln\left(p_{ijk}/q_{ijk}\right) - \sum_{i=1}^{m}\sum_{k=1}^{c}\alpha_{ik}\left(n_{i\cdot k} - \sum_{j=1}^{g} n_{ij\cdot}p_{ijk}\right) - \sum_{i=1}^{m}\sum_{j=1}^{g}\gamma_{ij}\left(\sum_{k=1}^{c} p_{ijk} - 1\right),$$

where $\boldsymbol{\alpha}$ is the Lagrange multiplier for constraint (3.4) and $\boldsymbol{\gamma}$ is the Lagrange multiplier for constraint (3.5). The solution of the first-order condition leads to the following expression for the conditional probabilities

$$\widehat{p}_{ijk} = \frac{q_{ijk}\exp\left(\widehat{\alpha}_{ik}n_{ij\cdot}\right)}{\sum_{k=1}^{c} q_{ijk}\exp\left(\widehat{\alpha}_{ik}n_{ij\cdot}\right)} \ . \tag{3.6}$$

In general, this solution does not have a closed-form expression, and the optimal values of the unknown parameters must be numerically determined.

If non-sample information exists about the conditional probabilities, $p_{ijk}$, this information can be introduced through the reference weights, $q_{ijk}$. As noted above, if non-sample information concerning the unknown conditional probabilities does not exist, then we use the uniform reference weights, $q_{ijk} = c^{-1}$, where $c$ is the number of candidates. In the case of uniform reference weights, the criterion function is

$$\arg\min_{p_{ijk}} \sum_{i=1}^{m}\sum_{j=1}^{g}\sum_{k=1}^{c} p_{ijk}\ln(p_{ijk}) \ , \tag{3.7}$$

and the solution is

$$\widehat{p}_{ijk} = \frac{\exp\left(\widehat{\alpha}_{ik}n_{ij\cdot}\right)}{\sum_{k=1}^{c}\exp\left(\widehat{\alpha}_{ik}n_{ij\cdot}\right)} \ . \tag{3.8}$$

The formulation (3.8) above avoids introducing any additional information that is not explicitly contained in the data. If the data constraint is omitted, uniform $p_{ijk}$ estimates result and provide the maximum entropy solution. Alternatively, in the context of (2.1) or (2.3), the function (3.7) is minimized subject to constraints imposed by the data and the additivity condition. An illustration of the solution process is given in Appendix B.

## 4. Empirical Results

Using election returns from the race for the 5th congressional seat in Louisiana in the 2000 election year, we will now demonstrate that our information theoretic methods may be used in reasoning through difficult VRA cases. In November

2000, Republican incumbent John Cooksey won the 5th district with 69% of the vote, against three Democrats who took 24%, 4%, and 3%, respectively. The district was about 65% white and 33% black, with very small Asian, Hispanic and mixed-race populations. We know what proportion of the vote each candidate received and the racial proportions of the electorate, and we would like to determine how the vote for a given candidate broke down on racial lines. If blacks voted overwhelmingly for one candidate while whites voted overwhelmingly for another candidate, then there would be evidence of racially polarized voting, and the judge could order the drawing of new district lines, fundamentally altering and essentially determining who would be elected.

Table 2: Precinct-Level Results from Information Theoretic Model. Louisiana's 5th CD

| **Ouachita Parish, Precinct 1–5** | | | | | | |
|---|---|---|---|---|---|---|
| | **R** | **D** | **I-1** | **I-2** | **A** | |
| **white** | 0.7580 | 0.1250 | 0.0008 | 0.0000 | 0.1161 | 1158 |
| **black** | 0.3539 | 0.2505 | 0.0959 | 0.0526 | 0.2470 | 222 |
| **other** | 0.2220 | 0.2116 | 0.1850 | 0.1702 | 0.2112 | 31 |
| | 963 | 207 | 28 | 17 | 196 | 1411 |
| **Rapides Parish, Precinct 38** | | | | | | |
| | **R** | **D** | **I-1** | **I-2** | **A** | |
| **white** | 0.4632 | 0.1241 | 0.0013 | 0.0000 | 0.4114 | 544 |
| **black** | 0.3133 | 0.2389 | 0.0929 | 0.0492 | 0.3057 | 112 |
| **other** | 0.2259 | 0.2142 | 0.1779 | 0.1571 | 0.2249 | 22 |
| | 292 | 99 | 15 | 9 | 263 | 678 |

**R**=Republican, **D**= Democrat, **I-1**= Independent 1, **I-2**= Independent 2, **A**= Abstention.

### 4.1 Uniform reference weights

We begin our analysis using non-informative, uniform $q_{ijk}$ reference distribution weights. To obtain a result for the entire district, one first computes the conditional probabilities, $\widehat{p}_{ijk}$ for each precinct $i$ individually, and then, given the objective, one may combine these separate precinct estimates into an overall estimate. Table 2 shows two representative precincts. Ouachita parish's precinct 1–5 had 1,158 white voters, 222 black voters, and 31 voters who were neither black nor white. Cooksey received 963 votes there, as against 207, 28, and 17 for his Democratic rivals, and 196 ballots cast in that precinct featured no valid vote

for the US House contest. All of the known information is thus displayed in the margins in Table 2. The estimates from our model are displayed in the interior cells of the table. The $p_{ijk}$ estimates in Table 2 are not adjusted to respect the constraint that the $N_{ijk}$ values must be integers. Except for small values of $N_{i\cdot\cdot}$, this oversight is inconsequential. Below, the same information is shown for a second precinct in the data set. We can observe from the marginals that these two particular precincts are similar in being largely white and heavily Republican, but we can also note by comparing the conditional probabilities that the estimated white voting rates differed quite substantially. In 2000, the Fifth Congressional District spanned 752 precincts in 20 parishes, and across these precincts there was quite a bit of variability in the number of voters in each racial group, the vote breakdown for the various candidates, and, relatedly, the resulting estimates from our model.

It is often necessary in VRA cases to be able to examine individual precincts rather than being presented with a single summary estimate for the entire district. Fortunately, our methods allow estimation of the $p_i$ for a single precinct based only on the data from that precinct thus allowing us to pin down and distinguish behavior among the various precincts.

### 4.2 Non-uniform reference weights

The results presented above were based on uniform reference weights, and thus did not incorporate any out-of-sample information. However, one could also choose reference weights that reflect additional information about the underlying process. In the context of elections, three obvious sources of potentially useful information are historical returns (particularly in those cases where electoral districts have not been redrawn), contemporaneous results in other contests, and exit polls. Historical returns have the feature that they are likely to be available at the same level of aggregation (the precinct, say), and so they are incompletely informative about the quantities of interest in precisely the same way as the contemporary returns. Because American ballots nearly always feature multiple contests, the available information for a given geographic unit such as a precinct usually includes returns for a number of races determined simultaneously. Finally, contemporary opinion polls offer an alternative (or additional) source of information, and, depending on the questions asked, may constitute estimates of the actual quantities of interest (the conditional probabilities). On the other hand, they are very unlikely to be available at the lowest level of aggregation. Instead, an opinion poll taken across a whole congressional district or state might be incorporated into a set of weights used as a basis for computing estimates in every precinct.

Table 3: Precinct-Level Cross-entropy. Louisiana's 5th CD

| **Cross-Entropy Weights** | | | | | |
|---|---|---|---|---|---|
| | **R** | **D-1** | **D-2** | **D-3** | **A** |
| **white** | 0.48 | 0.20 | 0.01 | 0.01 | 0.30 |
| **black** | 0.24 | 0.34 | 0.01 | 0.01 | 0.40 |
| **other** | 0.24 | 0.24 | 0.01 | 0.01 | 0.50 |
| **Cross-Entropy Estimates** | | | | | |
| **Precinct 1** | | | | | |
| | **R** | **D-1** | **D-2** | **D-3** | **A** |
| **white** | 0.7693 | 0.1082 | 0.0214 | 0.0122 | 0.0889 |
| **black** | 0.2907 | 0.3344 | 0.0128 | 0.0115 | 0.3506 |
| **other** | 0.2472 | 0.2401 | 0.0104 | 0.0102 | 0.4922 |
| **Precinct 2** | | | | | |
| | **R** | **D-1** | **D-2** | **D-3** | **A** |
| **white** | 0.4768 | 0.1095 | 0.0246 | 0.0139 | 0.3752 |
| **black** | 0.2442 | 0.3060 | 0.0123 | 0.0109 | 0.4267 |
| **other** | 0.2401 | 0.2344 | 0.0104 | 0.0101 | 0.5049 |

**R**=Republican, **D-1**= Democrat 1, **D-2**= Democrat 2, **D-3**= Democrat 3, **A**= Abstention.

To modify the reference weights to exploit additional information, one can follow Gokhale and Kullback (1978) and invoke the cross entropy principle, $\sum_{ijk} p_{ijk} \ln(p_{ijk}/q_{ijk})$, and the reference distribution information, $q_{ijk}$, as in (3.3). For our data, we consider some historical information. In 1998, Cooksey won the seat uncontested. In 1996, the first year those particular boundaries were in place, the seat was open, and Cooksey and Democrat Francis Thompson advanced from the primary (held in September that year) to a November run-off, which Cooksey won 58% to 42%. Precinct data from 1996 could thus be used to estimate the probabilities of voting for each candidate, conditional on race. In turn, one might use these 1996 estimates to generate reference weights for analysis of the 2000 returns. Louisiana and a few other southern states, because of past Court rulings stemming from Voting Rights cases, are unique in reporting not just racial composition of the electorate, but racial registration and turnout data as well. Hence, in this instance, an additional piece of information available for each precinct is the racial composition of the registered voter population. For instance, the 1158 white voters in the Ouachita precinct shown in Table 5 were about 50% registered Republicans and 33% registered Democrats; by contrast, in

the Rapides precinct, white voters were about 35% Republican and 50% Democrat. One might model voting rates in the U.S. House race (or any other race) using these data on composition of the voter pool.

Suppose, then, that on the basis of supplemental information, one chose identical reference weights for these two precincts, shown in Table 2. The estimates we obtain using the reference or cross-entropy weights shown in the top of Table 3 (rather than the uniform reference weights) are shown at the bottom of Table 3. Despite the very different reference weights used to obtain the results in Tables 2 and 3, the final estimates, themselves, are somewhat similar, but sport some significant differences. Most notably for the black group, our new priors on Democratic leanings shift the Republican-Democratic balance dramatically from the results previously obtained with uniform priors. In general, one can explore how the estimated conditional probabilities respond to a range of reference weights as a means of investigating the range of possible substantive conclusions consistent with the data.

## 5. Discussion

We have presented an information theoretic approach as a framework for reasoning in voting rights cases where the data are in aggregate form and the basic corresponding recovery problem is a pure ill-posed inverse problem. Using a maximum entropy estimation criterion and aggregate election data formulated in pure inverse problem form, we have demonstrated the information theoretic method using actual election returns. Throughout, our emphasis has been on formulating an information recovery procedure predicated on using only the information available to the researcher and minimizing the use of creative assumptions to convert an ill-posed problem into a well-posed one that can be solved by conventional means.

All formulations seeking to process and recover information from observed aggregate election data must include some assumptions. In this paper, we have attempted to make use of Occam's razor, the logical principle that a problem or model should be stated and analyzed in its most basic and simplest terms and minimize the use of unsupported assumptions. Our emphasis has been on developing a tractable way to extract information from aggregate data when data restrictions mean that it must be modelled as an ill-posed inverse problem. In Section 4.2, we demonstrated how prior beliefs incorporated into a reference distribution may be used to augment the sample information. In this framework, one can incorporate behavioral models and employ recovery procedures consistent with the underlying outcome data. This conceptual basis can be extended to allow the investigator to employ covariates that may be used to condition the unknown probabilities (Judge, Miller and Cho, 2004). In the information recovery method

used in this study, the assumptions are minimal and clear, and the solution satisfies the simple and attractive multiplicity principle.

**Acknowledgements**

Thanks to Brian Gaines, Ken Benoit, Jake Bowers, Bruce Cain, Gary King, Jim Kuklinski, Joanne Lee, Ken McCue, Rogerio de Mattos, Peter Nardulli, Ori Rosen, and Kevin Quinn for helpful and substantive comments.

## Appendix A. Computational aspects: obtaining a solution

### A.1 Computational aspects

There are two basic methods of approaching the computational aspect of this problem. One is to write a program to solve a constrained minimization problem. The other is to write a program to solve an unconstrained minimization problem. For our particular problem, the constraints for our constrained optimization problem are shown in (3.4) and (3.5). Given this constrained optimization setup, one could use a software package such as GAUSS or GAMS that can perform constrained optimization to obtain results. Alternatively, one could use the concentrated objective function

$$F(\lambda) = \sum_{k=1}^{c} n_{\cdot k}\boldsymbol{\lambda}_k + \sum_{j=1}^{g} \ln\left[\sum_{k=1}^{c} \exp(-\boldsymbol{\lambda}_k n_{j\cdot})\right] \tag{A.1}$$

in any program that contains an optimization routine (constrained or unconstrained). Splus and R, for instance, have general optimization routines for unconstrained optimization. To obtain results for an entire district, one needs to perform this optimization for every observation (e.g. precinct) in the data set. So the results need to be combined to obtain a single estimate for an entire district. Sample code is available from the authors upon request.

A notable computational advantage of the approach to this problem based on the Kullback-Leibler cross-entropy (3.6) or the Shannon entropy functional (3.8) is that two of the main restrictions on the weights can be seamlessly and easily built into the computational process. First of all, the inequality constraint that requires nonnegativity ($p_i \geq 0$), is clearly satisfied since $p(\lambda)$ is representable as an exponential function of the parameters and data. This function is inherently nonnegative. Second, given that nonnegativity is automatically imposed, the additivity restriction can be enforced by the functional form, where the exponential function is divided by a normalizing sum of the exponentials. Accordingly, except for the moment/estimating equation itself, which still needs to be enforced as a constraint, all of the other constraints are seamlessly integrated without any additional requirements of the optimizing algorithm.

## Appendix B. An illustration of the solution process

Consider an election with 3 candidates, identified as "1", "2", and "3". Suppose that $N$ votes are cast and that the individual votes are recorded as $v_i = k$, where $i = 1, \ldots, N$ indicates the voter and $k = 1, 2, 3$ indicates voter $i$'s vote choice. To keep the example simple, suppose that the results from the election are tabulated as the total sum of the votes ($\sum_i v_i$), and only an average over all of the votes ($\overline{v} = \sum_i v_i/N$) is reported. Given this information, how would one estimate what proportions of the vote, $p_k$, each candidate received?

With sufficiently large $N$, the number of distributions supported on set $S = \{1, 2, 3\}$ for each possible mean, $\overline{v}$, is large enough that it does no harm to ignore the discrete nature of the problem and treat it as infinite. Then, this inverse problem is clearly ill-posed: we have three unknown probabilities, $p_1$, $p_2$, $p_3$, but only two pieces of information. The probabilities sum to 1,

$$\sum_{k=1}^{3} p_k = p_1 + p_2 + p_3 = 1 \ ,$$

and the average vote score is

$$\sum_{k=1}^{3} p_k \, x_k = \overline{v} \ .$$

where $x_k = k$ for $k = 1, 2, 3$.

In order to solve this problem by using all the information we have while also minimizing reliance on information we do not have, we select the probabilities that maximize the Shannon entropy function,

$$H(p) = -\sum_{k=1}^{3} p_k \, \ln(p_k) \ ,$$

subject to two constraints,

$$\sum_{k=1}^{3} p_k \, x_k = \overline{v} \ , \tag{B.1}$$

and

$$\sum_{k=1}^{3} p_k = 1 \ . \tag{B.2}$$

with $p_k \geq 0$.

Since this is a basic maximization problem subject to two constraints, we use the method of Lagrange multipliers to arrive at a solution. The Lagrangian for this problem is

$$L(\boldsymbol{p}, \alpha, \gamma) = -\sum_{k=1}^{3} p_k \, \ln(p_k) + \alpha \left( \overline{v} - \sum_{k=1}^{3} p_k \, x_k \right) + \gamma \left( 1 - \sum_{k=1}^{3} p_k \right) , \qquad \text{(B.3)}$$

where $\alpha$ and $\gamma$ are the Lagrange multipliers for constraints (B.1) and (B.2) respectively. The first-order condition is

$$\frac{\partial L}{\partial p_k} = -\ln(p_k) - 1 - \alpha \, x_k - \gamma = 0.$$

Solving this yields

$$\begin{aligned} \ln(p_k) &= -\alpha \, x_k - 1 - \gamma \\ p_k &= \exp\{-\alpha \, x_k - 1 - \gamma\}. \end{aligned}$$

We can express the intermediate solution as a function of the Lagrange multipliers,

$$\widehat{p}_k = \frac{\exp\left(-\widehat{\alpha} \, x_k\right)}{\sum_{k=1}^{3} \exp\left(-\widehat{\alpha} \, x_k\right)} , \qquad \text{(B.4)}$$

where $\widehat{\alpha}$ is the optimal Lagrange multiplier for constraint (B.1).

We can substitute (B.4) back into the Lagrange equation (B.3) to obtain a concentrated objective function

$$F(\alpha) = \alpha \, \overline{v} + \ln \left[ \sum_{k=1}^{3} \exp\left(-\alpha \, x_k\right) \right] . \qquad \text{(B.5)}$$

Equation (B.5) is strictly convex in $\alpha$. We can find the optimal value of the Lagrange multiplier, $\alpha$, by minimizing $F(\alpha)$. We can then use this value of the Lagrange multiplier, $\widehat{\alpha}$, in Equation (B.4) to obtain an estimate of the vote share. This simple problem is a variant of Jaynes's famous dice problem in which one must assign probabilities to the six faces of a die based on the observed average outcome of $N$ rolls. In our case, we have three unknown probabilities and only two pieces of available information.

Table 4: Estimated Conditional Probabilities

| $\overline{v}$ | $\widehat{p}_1$ | $\widehat{p}_2$ | $\widehat{p}_3$ | $H(\widehat{\boldsymbol{p}})$ |
|---|---|---|---|---|
| 1.5 | 0.6162 | 0.2676 | 0.1162 | 0.9012 |
| 2.0 | 0.3333 | 0.3333 | 0.3333 | 1.0986 |
| 2.5 | 0.1162 | 0.2676 | 0.6162 | 0.9012 |

Although this is a simple, unrealistic example, it is nonetheless representative of a large class of problems in voting-behavior, where information is limited to aggregate returns. It exemplifies the constraint-based problems developed in Sections 2 and 3 since there are three unknowns (any two of which completely determine the third), but there is only one data point, making the problem ill-posed. The constraint involving $\overline{v}$ is the only aggregate information available. Given $\overline{v}$ and the structure of the problem, for a large $N$, the number of ways that the constraints (B.1) and (B.2) can be satisfied is practically infinite. Using conventional rules of logic, there is no way to solve this problem uniquely. By utilizing the information theoretic criteria, however, we are led to an optimal constraint-based formulation that yields a solution that is consistent with what we know (i.e. the data). This solution has some attractive characteristics, not least of which is that it makes minimal use of assumptions or information to solve the estimation problem at hand.

## References

Baggerly, K. (1998). Empirical likelihood as a goodness of fit measure. *Biometrika* **85**, 535-547.

Burg, J. P. (1967). Maximum entropy spectral analysis. *Proceedings of the 37th Meeting of the Society for Exploration Geophysicists*. Reprinted in *Modern Spectrum Analysis* (Edited by D. G. Childers), 34-39. IEEE Press.

Cressie, N., Read, T. R. C. (1984). Multinomial goodness of fit tests. *Journal of the Royal Statistical Society, Series B* **46**, 446-464.

Gokhale, D. and Kullback, S. (1978). *The Information in Contingency Tables*. Marcel Dekker.

Good, I. J. (1963). Maximum entropy for hypothesis formulation, especially for multi-dimensional contingency tables. *Annals of Mathematical Statistics* **34**, 911-934.

Goodman, L. A. (1953). Ecological regressions and behavior of individuals. *American Sociological Review* **18**, 663-669.

Goodman. L. A. (1959). Some alternatives to ecological correlation. *American Journal of Sociology* **64**, 610-625.

Jaynes, E. T. (1957). Information theory and statistical mechanics ii. *Physical Review* **108**, 171-190.

Judge, G. G., Miller, D. J. and Cho, W. K. T. (2004). An information theoretic approach to ecological estimation and inference. In *Ecological Inference: New Methodological Strategies* (Edited by Gary King, Ori Rosen and Martin Tanner), 162-187. Cambridge University Press.

Jumarie, G. (1990). *Relative Information: Theories and Applications*. Springer-Verlag.

King, G. (1997). *A Solution to the Ecological Inference Problem*. Princeton University Press.

Kullback, S. (1958). *Information Theory and Statistics*. John Wiley and Sons.

Mittelhammer, R., Judge, G. G. and Miller, D. J. (2000). *Econometric Foundations*. Cambridge University Press.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single function. *Biometrika* **75**, 237-249.

Owen, A. B. (1990). Empirical likelihood ratio confidence regions. *Annals of Statistics* **18**, 90-120.

Read, T. R. C. and Cressie, N. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer-Verlag.

Shannon, C. E. (1948). A mathematical theory of communications. *The Bell System Technical Journal*, **27**, 379-423, 623-656.

Shannon, C. E. (1949). Communication in the presence of noise. *Proceedings of the Instutute of Radio Engineers* **37**, 10-21.

Wakefield, J. (2004). Ecological inference for 2 $\times$ 2 tables. *Journal of the Royal Statistical Society, Series A*, **167**, 385-445.

Received November 21, 2006; accepted March 7, 2007.

Wendy K. Tam Cho
Departments of Political Science and Statistics
University of Illinois at Urbana-Champaign
361 Lincoln Hall
702 S. Wright St.
Urbana, IL 61801–3696. USA
wendycho@uiuc.edu

George G. Judge
Graduate School
University of California at Berkeley
207 Giannini Hall
University of California
Berkeley, CA 94720–3310 USA
judge@are.berkeley.edu