Author's personal copy

Available online at www.sciencedirect.com

ScienceDirect

Economics Letters 97 (2007) 201–207

economics letters

www.elsevier.com/locate/econbase

Information theoretic solutions for correlated bivariate processes ☆

Wendy K. Tam Cho [a,*], George G. Judge [b]

[a] Departments of Political Science and Statistics and Senior Research Scientist at the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, USA

[b] University of California at Berkeley, 207 Giannini Hall, University of California, Berkeley, CA 94720-3310, USA

Received 18 August 2006; received in revised form 1 March 2007; accepted 15 March 2007
Available online 6 July 2007

Abstract

In a bivariate context, we consider ill-posed inverse problems with incomplete theoretical and data information. We demonstrate the use of information theoretic methods for information recovery for a range of under-identified choice problems with more unknowns than data points.

© 2007 Elsevier B.V. All rights reserved.

Keywords: Information theoretic methods; Maximum entropy; Multidimensional contingency tables

JEL classification: C1; C10; C24

1. Introduction

In this paper we consider the problem of information recovery in the case of bivariate discrete distributions when data exist only in the form of marginal totals. Conventional procedures for this problem are typically based on strong assumptions involving finitely parameterized specifications. The substantive theories that motivate these models and estimation rules rarely justify such restrictions. Within this context, we demonstrate an alternative approach to information recovery that has the following characteristics: i) the observables exist only in the form of macro or aggregate outcome data; ii) the

☆ Thanks to Peter Danaher, Bruce Hardie, and Douglas Miller for helpful comments, and to Peter Danaher and Bruce Hardie for providing a spreadsheet of their analysis.

* Corresponding author. 361 Lincoln Hall, 702 S. Wright St., Urbana, IL 60801-3696, USA. Tel.: +1 217 333 9588; fax: +1 217 244 5712.

E-mail addresses: wendycho@uiuc.edu (W.K.T. Cho), judge@are.berkeley.edu (G.G. Judge).

0165-1765/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.econlet.2007.03.020

Author's personal copy

corresponding micro data in many cases is not only unobserved but unobservable; iii) information concerning the underlying data sampling process may be partial, incomplete, and insufficient to identify a unique family of feasible parametric models; and iv) consistent with these limited theoretical and data situations, information recovery possibilities exist only in the form of an ill-posed pure inverse problem. To cope with these types of situations, in the Sections ahead, we specify the binomial correlated processes as an ill-posed inverse problem and demonstrate how information theoretic methods may be used to provide a solution basis to illustrate the information recovery possibilities. We illustrate these methods with two examples, one on consumer choices in the purchase of bacon and eggs and one on voting and candidate choice.

2. The bivariate model

In order to examine this problem, we first provide a notational base. Designate the outcomes of the random variables as Y_j=0, 1, 2, …, J and Y_k=0, 1, 2, …, K. The observed information is then $N_k = \sum_{j=1}^{J} N_{jk}$, $N_j = \sum_{k=1}^{K} N_{jk}$, and $N = \sum_{j=1}^{J} \sum_{k=1}^{K} N_{jk}$. Our objective is to formulate a pure inverse problem that will permit us to recover estimates of N_{jk}, the unobserved outcomes, by using only the observed aggregate marginal data. The unobserved choice or behavior outcomes may be expressed in terms of the observed row and column proportions, n_k=N_k/N and n_j=N_j/N, and the proportion of consumers in each category p_{jk}=N_{jk}/N_j=n_{jk}/n_j, where $\sum_{k=1}^{K} p_{jk} = 1$. In this context, p_{jk} is the conditional probability and j is the conditioning index.

2.1. Modeling behavior as an ill-posed pure inverse problem

If the conditional probabilities, p_{jk}, in the interior cells of Table 2 were known, we could derive the unknown behavioral response or choices as N_{jk}=$p_{jk}N_j$. However, because the conditional probabilities in many problems are unobserved and not accessible by direct measurement, we face an inverse problem where indirect, partial, and incomplete aggregate data must be used to recover the unknown conditional probabilities. Some structure is provided by the realization that the conditional probabilities, p_{jk}, must satisfy the additivity condition, $\sum_{k=1}^{K} p_{jk} = 1$, and the column sum conditions, $\sum_{j=1}^{J} p_{jk}N_j = N_k$. The column sum conditions give us the relationship

$$n_k = \sum_{j=1}^{J} n_j p_{jk}, \tag{1}$$

for k=1,…, K. To formalize our notation, we let $\mathbf{x}$=$(n_1\ n_2\ \cdots\ n_J)'$ represent the $(J\times 1)$ vector of proportions, j=1,…, J, and let $\boldsymbol{y}$=$(n_1\ n_2\ \cdots\ n_K)'$ represent the $(K\times 1)$ sample outcome vector of vote proportions for k=1,…, K. The relationship among the observed marginal proportions and unknown conditional probabilities may be written as

$$\boldsymbol{y}' = \mathbf{x}'\mathbf{P}, \tag{2}$$

where the component $\mathbf{P}$=$(\boldsymbol{p}_1\ \boldsymbol{p}_2\ \cdots\ \boldsymbol{p}_K)$ is an unknown and unobservable $(J\times K)$ matrix of conditional probabilities, and $\boldsymbol{p}_k$=$(p_{1k}\, p_{2k}\ \cdots\ p_{Jk})'$ is the $(J\times 1)$ vector of conditional probabilities associated with the kth group.

Author's personal copy

The formulation in (2), connecting the unknown and unobservable proportions, is in the form of a pure ill-posed inverse problem, 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 $\boldsymbol{p}$ is an *unknown* high dimensional parameter vector. The inverse problem is to recover the unobservable $\boldsymbol{p}$'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, using traditional rules of logic, insufficient sample information exists to solve the problem using traditional rules of logic.

2.2. Information theoretic formulation and solution

To implement the model of behavior introduced in Section 2, we must determine how to represent the data and how to choose the criterion or objective function. The representation of the data is discussed in connection with (1) and (2). Because of the ill-posed nature of the inverse problem (2), traditional estimation methods cannot be used to recover the unknown p_{jk}. One possibility is to introduce structure in the way of creative assumptions, parametric or otherwise. To avoid adding extraneous information that the researcher usually does not possess, we make use of the information theory contributions of Claude Shannon (1948, 1949), in choosing a criterion function. 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

$$-H(\boldsymbol{Y}) = p_1 \log p_1 + \ldots + p_n \log p_n = -\sum_i p_i \ln(p_i). \tag{3}$$

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 data constraints, the distribution that maximizes entropy.[1]

Under the Shannon and Jaynes maximum entropy estimation criterion, the pure inverse model (2) may be formulated as

$$\arg\min_{p_{jk}} \sum_{j=1}^{J} \sum_{k=1}^{K} p_{jk} \ln(p_{jk}), \tag{5}$$

[1] The MaxEnt principle or criterion, $-\sum_i p_i \ln(p_i)$, is a member of the Cressie–Read (Cressie and Read, 1984; Read and Cressie, 1988) family of minimum divergence distance measures. The Cressie–Read power-divergence (CR) statistic (Cressie and Read, 1984; Read and Cressie, 1988; Baggerly, 1998)

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

provides a family of distance or discrepancy measure between $\boldsymbol{p}$ (i.e., the conditional probabilities in our problem) and a set of reference weights $\boldsymbol{q}$. When $\lambda=0$, and the reference distribution, q_i, is uniform, the Cressie-Read distance measure yields the Shannon/Jaynes entropy criterion, $-\sum_i p_i \ln(p_i)$. For a more complete discussion of the CR statistic and corresponding family of criterion functions, see Mittelhammer et al. (2000). For a discussion of the entropy principle, see Golan et al. (1996).

Author's personal copy

subject to the column–sum condition,

$$n_{\cdot k} = \sum_{j=1}^{J} n_j p_{jk}, \tag{6}$$

and the additivity condition,

$$\sum_{k=1}^{K} p_{jk} = 1 \quad \forall j. \tag{7}$$

In this way, the problem is stated as a constrained minimization problem that minimizes the distance between the estimated p_i and q_i, a uniform reference distribution. Depending on the external knowledge base, other fixed or random q_i may serve as the reference distribution. Note that the statement of the pure inverse problem, in an extremum context, involves three components: the distance measure (5), the data constraint (2) in the form of (6), and the additivity condition (7).

The Lagrangian function for the constrained minimization problem expressed in (5), (6) and (7) is

$$L(\boldsymbol{p}, \boldsymbol{q}, \lambda, \boldsymbol{a}, \gamma) = \sum_{j=1}^{J} \sum_{k=1}^{K} p_{jk} \ln(p_{jk}) - \sum_{k=1}^{K} \alpha_k \left(n_k - \sum_{j=1}^{J} n_j . p_{jk} \right) - \sum_{j=1}^{J} \gamma_j \left(\sum_{k=1}^{K} p_{jk} - 1 \right), \tag{8}$$

where $\boldsymbol{\alpha}$ is the Lagrange multiplier for constraint (6) and $\boldsymbol{\gamma}$ is the Lagrange multiplier for constraint (7). The solution of the first-order condition leads to the following expression for the conditional probabilities:

$$\hat{p}_{jk} = \frac{\exp(\hat{\alpha} k n_j)}{\sum_{k=1}^{K} \exp(\hat{\alpha} k n_j)}. \tag{9}$$

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

3. Applications

We use two examples to illustrate the use of the maximum entropy approach for information recovery. The first is an application from economics and relates to consumer behavior in the purchase of bacon and eggs. The second is an application from political science and concerns voter behavior and candidate choice.

3.1. Bacon and eggs

In a recent paper, Danaher and Hardie (2005) consider a bivariate bacon and eggs problem. Their data are shown in Table 1. Using the data in the marginals of Table 1, the entropy criterion, and the solution basis developed in Section 2.2, our results are presented in Table 2. The conditional

Author's personal copy

Table 1
Observed bivariate distribution of the number of times bacon and eggs were purchased on four consecutive shopping trips

Bacon	Eggs					Total
	0	1	2	3	4	
0	254 (0.5907)	115 (0.2674)	42 (0.0977)	13 (0.0302)	6 (0.0140)	430
1	34 (0.3953)	29 (0.3372)	16 (0.1860)	6 (0.0698)	1 (0.0116)	86
2	8 (0.3478)	8 (0.3478)	3 (0.1304)	3 (0.1304)	1 (0.0435)	23
3	0 (0.0000)	0 (0.0000)	4 (0.6667)	1 (0.1667)	1 (0.1667)	6
4	1 (0.3333)	1 (0.3333)	1 (0.3333)	0 (0.0000)	0 (0.0000)	3
Total	297	153	66	23	9	548

probabilities are listed in parentheses. The observable aggregate data are reflected in row and column sums, and the unknown or unobservable data have the respective conditional probabilities, p_{jk} in the interior cells of the table (see Good, 1963; Gokhale and Kullback, 1978). There is a fairly close fit between the actual data (Table 1) and the information theoretic estimates (Table 2). The fit tends to decline with the smaller values that are prone to occur at the edges of the table but, in general, the recovery pattern is impressive. The correlation between the observed and the estimated values for the first three rows is 0.999, 0.970, and 0.861, respectively.

3.2. Voting rights

Another example can be found in the analysis of election data (see Cho and Judge, 2007). In the U.S., we are subject to the secret ballot, and so while we may know how many votes a particular

Table 2
Empirical likelihood estimates

Bacon	Eggs					Total
	0	1	2	3	4	
0	262.34 (0.6101)	122.48 (0.2849)	40.48 (0.0941)	4.65 (0.0108)	0.01 (0.0000)	429.96
1	27.36 (0.3183)	23.50 (0.2734)	18.83 (0.2190)	12.21 (0.1420)	4.07 (0.0473)	85.97
2	5.38 (0.2342)	5.16 (0.2246)	4.87 (0.2120)	4.33 (0.1885)	3.23 (0.1406)	22.97
3	1.25 (0.2090)	1.24 (0.2074)	1.22 (0.2040)	1.18 (0.1973)	1.09 (0.1823)	5.98
4	0.61 (0.2047)	0.61 (0.2047)	0.60 (0.2013)	0.59 (0.1980)	0.57 (0.1913)	2.98
Total	296.94	152.99	66.00	22.96	8.97	

Author's personal copy

Table 3
Precinct-level results from information theoretic model

	Republican	Democrat	Independent 1	Independent 2	Abstention	Total
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

Louisiana's 5th CD.

candidate received in a particular precinct, we do not know how particular individuals or groups of individuals voted. We can obtain, for example, racial demographic information for a precinct but not the vote preferences of the racial groups. Accordingly, we know what proportion of the vote each candidate received and the racial proportions of the electorate, but we would like to determine how the different racial groups voted. The results of an analysis on racial vote preferences are often pivotal in a judge's decision regarding whether district lines must be redrawn.

Consider the election data shown in Table 3. Because of the secret ballot, we know the marginal row and column totals, but not the values in the interior cells, p_{ij}. All of the known information is thus displayed in the margins. In this particular election, there were four candidates. The estimated conditional probabilities are displayed in the interior cells of Table 3. In this case, as in virtually all cases found in practice, we do not know if our estimated conditional probabilities match the empirical reality, since the empirical outcomes are masked by the secret ballot or data that are partial and incomplete. However, the information recovered via the information theoretic approach provides one valuable basis for decision making and choice.

4. Summary

The information theoretic approach provides an appealing basis for information recovery in the context of pure ill-posed inverse problems. If the unknowns of the problem are unobservable and only marginal or aggregate data totals are available, information theoretic methods provide a way to recover information from indirect observations via a solution to an ill-posed inverse problem. The information theoretic approach avoids a fully parametric/structural approach and proceeds instead with a minimum number of assumptions. This is in accord with the logical principle of Occam's razor. In ill-posed inverse or under-identified problems, even after we limit the solution set to estimates consistent with data constraints such as (6) and (7), there are an infinite number of remaining possible solutions. Using the MaxEnt criterion, the solution chosen from among all of the possible solutions is the one that could happen in the most likely or greatest number of ways.

If in cases such as those involving economic behavior where only macro data is available, MaxEnt conditional probabilities provide null hypotheses that may be checked by way of an experiment or survey. Information theoretic methods are easy to implement and computationally simple. If one has theoretical information concerning the conditional probabilities such as distributional symmetry or information from a previous sample, it is simple to incorporate this information in the reference distribution and to modify the model accordingly.

Author's personal copy

References

Baggerly, K., 1998. Empirical likelihood as a goodness of fit measure. Biometrika 85 (3), 535–547.

Cho, W.K.T., Judge, G.G., 2007. Recovering vote choice from partial incomplete data. Journal of Data Science (Forthcoming).

Cressie, N., Read, T.R.C., 1984. Multinomial goodness of fit tests. Journal of the Royal Statistical Society. Series B 46, 440–464.

Danaher, P.J., Hardie, B.G.S., 2005. Bacon with your eggs? Applications of a new bivariate beta-binomial distribution. The American Statistician 59 (4), 282–286.

Gokhale, D., Kullback, S., 1978. The Information in Contingency Tables. Marcel Dekker, New York.

Golan, A., Judge, G.G., Miller, D.J., 1996. Maximum Entropy Econometrics: Robust Estimation with Limited Data. John Wiley and Sons, New York.

Good, I.J., 1963. Maximum entropy for hypothesis formation, especially for multidimensional contingency tables. The Annals of Mathematical Statistics 34 (3), 911–934.

Jaynes, E.T., 1957. Information theory and statistical mechanics II. Physical Review 108, 171–190.

Mittelhammer, R., Judge, G.G., Miller, D.J., 2000. Econometric Foundations. Cambridge University Press, New York.

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

Shannon, C.E. 1948. A Mathematical Theory of Communication. The Bell System Technical Journal 27:379–423, 623–656.

Shannon, C.E., 1949. Communication in the presence of noise. Proceedings of the IRE 37, 10–21.