

Multi Level Categorical Data Fusion Using Partially Fused Data

Zvi Gilula Robert McCulloch *

December 18, 2011

Abstract

Data fusion poses challenging methodological issues for inferring the joint distribution of two random variables when the information available is mainly confined to the marginal distributions. When the variables are categorical, the challenges are even more severe. Applications of categorical data fusion are of top importance in marketing, especially in advertising. A great deal of categorical data fusion methods are confined to binary variables. In this paper we develop an innovative approach to categorical data fusion that extends previous methodologies and applies to categorical variables with any number of levels. We introduce a new concept of “evident dependence” that describes a variety of patterns of joint distributions given the marginals. Using information from partially fused data, our method smoothly accommodates a non-trivial Bayesian approach based on mixtures of joint distributions constructed using evident dependence. The approach is illustrated using data from the advertising industry.

KEY WORDS: evident dependence, mixture modeling, copulas.

*Zvi Gilula is Professor of Statistics, Department of Statistics, Hebrew University, msgilula@mssc.huji.ac.il. Robert E. McCulloch is Professor of Statistics, IROM Department, 1 University Station, B6500, Austin, TX 78712-1175, robert.mcculloch1@gmail.com.

Contents

1	Introduction	1
2	Constructing Evident Dependence Tables from Marginals	3
3	Detailed Description of the Proposed Approach	7
3.1	Framework for Categorical Data Fusion	7
3.2	Relaxing the Conditional Independence Assumption – The GMR Approach .	8
3.3	A New Fusion Approach for Multi Level Multinomials: Evident Dependence Mixture	10
3.4	Prior and Modeling Choices	11
3.5	Approximate and Exact Inference	12
4	Empirical Results	13
4.1	The Data	14
4.2	The Kullback-Leibler Divergence	15
4.3	Evident Dependence Fusion: A Sampling Experiment	16
5	Conclusion	21
6	Appendix: The GMR Approach as a Special Case of Evident Dependence	22

1 Introduction

In data fusion we seek to make inference about the joint distribution of a pair of variables even though we do not observe them together.

For example, one survey, taken by researchers studying our health, might ask respondents how much they smoke cigarettes (the variable “cig”). Another survey, taken by policy researchers, might ask people how they feel about banning smoking in public (the variable “ban”). We could wonder about the relation between the two variables cig and ban, but have no survey in which both questions are asked. In this case, we have information about the marginal distributions of cig and ban, but we have no observations from the joint distribution. In a typical survey setting, each question asks the respondent to choose from a small set of options so that the variables would be categorical.

We call the two variables, whose joint we seek, the “target variables”. The methods developed in this paper assume the each of the target variables is categorical. Our goal is the estimation of the two-way table corresponding to the joint distribution of the pair of target variables.

We cannot solve the data fusion problem without additional information or assumptions. (Kiesel & Rassler (2007)). If we assume the target variables are independent, we can obtain the joint from the marginals but we do not want to make this assumption. The assumption often used in practice is that the target variables are conditionally independent given a set of variables which are observed jointly with each of the variables in our target pair.

In our survey example, both surveys might include questions about the respondent such as the respondent’s age, income, and gender. We could then assume that cig and ban are independent given age, income, and gender.

We call the variables observed with each of the target variables the “common variables”. We have one data set with the first target variable and the common variables and another data set with the second target variable and the common variables.

While the conditional independence assumption can work remarkably well in practice, it may not be valid given the available set of common variables. In this paper we assume that the two data sets having one of the target variables and the common variables are large and that we have a third, relatively small data set, in which we observe the common variables and both of the target variables. Thus, we have a small amount of data where we do observe the two target variables together. We wish to combine information from all three data sets in a simple way to obtain an estimate of the table giving the joint distribution of the target

variables.

We combine the information in the three data sets using a model for dependence of the target variables given the common variables and a simple prior formulation which shrinks our inference towards the conditional independence solution without imposing it. Our model for conditional dependence is based a new approach to thinking about dependence in two-way tables which we dub “evident dependence”. In our examples we show that we can obtain better estimates by combining information in this way than we can by either assuming conditional independence or just using the information in the third data set where the target variables are observed jointly.

It is now quite easy and common to conduct a small survey. The methods of this paper allow researchers to combine information from large institutional surveys with smaller surveys more focused on the particular question of interest.

Gilula, McCulloch & Rossi (2006), (henceforth GMR) developed an approach for the same setup, but assumed that each of the target variables was binary (categorical with only two possible outcomes). The GMR approach does not extend in any obvious way to categorical variables with more than two outcomes. Kamakura & Wedel (1997) and Kamakura & Wedel (2000) proposed an elegant and innovative methodology dealing with higher dimensional categorical data fusion, but the models proposed there are far from parsimonious and are based on modeling the joint distribution of both the target variables and the common variables. GMR developed a ”direct” approach that does not require such weighty joint distributions. While the essential components of our approach are new to this paper, we share much of our basic framework with GMR and show that the GMR approach is actually a special case of the one developed here. Both GMR and this paper emphasize the application of data fusion ideas in the analysis of survey data collected for marketing research.

The term “data fusion” is often used to refer methods which construct a single “fused” data set with synthesized observations on the common variables and both target variables. This is typically done by taking an observation with one of the target variables and the common variables and then “matching” this observation with one from the data set having the other target variable. The matching is done by looking for an observation which has similar values of the common variables. We can then construct a single observation by augmenting the original observation with the value of the second target variable taken from the matched observation (see Kadane (1978) and Rogers (1984)). Rassler (2002) provides an excellent discussion of fusion methods. Rubin (1986) views the problem as a missing data problem. That is, the information on one target variable is “missing” from one of the data sets and

must be imputed using information from the other data sets. Note that in this paper, our goal is limited to the estimation of the table giving the joint distribution of the two target variables.

In Section 2 we present the concept of an evident dependent joint distribution. In Section 3 we describe our new approach based on conditional mixtures of evident dependence distributions. In Section 4 we illustrate the efficacy of our fusion approach using data coming from surveys on purchase behavior and media usage of close to 25,000 households in the UK. Section 5 concludes the paper with an overall summary and points out other areas (like copulas) where the proposed approach might be relevant and helpful. In the appendix we show that our approach nests that of GMR.

2 Constructing Evident Dependence Tables from Marginals

Let b and m be two categorical variables. The joint distribution of (b, m) may be given by the two-way table whose (i, j) entry p_{ij} is the probability b takes on its i^{th} possible level and m takes on its j^{th} possible level. In this section we describe our construction of evident dependence tables given the marginals of b and m .

For joint probabilities denoted by $\{p_{ij}\}$ and their corresponding marginals $\{p_{i.}\}$ and $\{p_{.j}\}$, the well-known Frechet bounds (see Tchen (1980)) on the joint probabilities given the marginals are

$$\max\{0, p_{i.} + p_{.j} - 1\} \leq p_{ij} \leq \min\{p_{i.}, p_{.j}\}.$$

A table $\{p_{ij}\}$ with entries on or near the boundaries reflects strong dependence (hence, *evident dependence*). If the upper bound is attained with $p_{ij} = p_{i.}$ for row i and column j then all other entries in row i must be zero, for all $j^* \neq j$. If the upper bound is attained with $p_{ij} = p_{.j}$ then all other entries in the j^{th} column must be zero for all $i^* \neq i$.

Our evident dependence algorithm seeks to use the largest possible p_{ij} elements (the upper bound) resulting in tables with zero entries. The algorithm leans on research on these bounds reported by Dobra & Feinberg (2001) and Dobra & Feinberg (2003) and on a recent paper Dall'Aglio & Bona (2011).

As discussed in Section 3.3, we wish to generate a variety of evident dependence tables. A particular table will correspond to a choice of a “rule” for each of the marginals. Each rule considers a vector of non-negative numbers and chooses a particular positive number from

the vector. The set of rules we will consider are specified by positive integers r . The rule chooses the element of the vector which is the r^{th} largest positive element. If there are an insufficient number of positive elements, the smallest positive element is chosen. So, if $r = 1$, the rule chooses the largest non-zero element. If $r = 2$, the second largest is chosen. Each evident dependence table is generated by first choosing a rule for the b marginal and a rule for the m marginal.

The algorithm is written below in a simple and common algorithmic language to allow users to both easily follow it and to be able to program it. The algorithm starts with a blank (all entries are zero) two-way table and then iteratively “fills in” elements of the table given the marginals and choice of rules. The quantity SUM is used in the algorithm to denote the remaining probability to be filled into the table. We start with SUM=1 for the blank table. Each time a probability is assigned to a cell in the table it is subtracted from SUM and from the marginals. Once SUM goes down to zero the Evidence Dependence table is complete. It is obvious that each time a cell is filled, all other entries in the row AND/OR the column defining that cell must either be zero for the row AND/OR the column or be a positive value that makes all entries in the underlying row and column add up to the corresponding marginals. As is exemplified, our algorithm guarantees that as many entries as possible attain the Frechet bounds and hence, results in joint distribution demonstrating strong dependence between the target variables b and m .

In stating our algorithm, we use the following notation. T denotes the table giving the joint distribution of b and m . The rows of T correspond to possible outcomes for b and the columns correspond to possible outcomes for m . Given a rule for b and vector of non-negative numbers P_b , we shall denote the chosen value by v_b and the index of the value by i_b . For example, if $P_b = (.2, .3, .5)$ and the b rule picks the second biggest element ($r = 2$), then $v_b = .3$ and $i_b = 2$. Thus, $v_b = P_b(i_b)$. Analogous definitions apply to the m rule.

Evident Dependence Algorithm

Initialize all elements of the table T to 0.

Initialize P_b to be the given marginal for b .

Initialize P_m to be the given marginal for m .

SUM = 1

while (SUM > 0)

 Apply the b rule to P_b giving the value v_b and index i_b .

 Apply the m rule to P_m giving the value v_m and index i_m .

 Let $A = \min(v_b, v_m)$

 Subtract A from the i_b element of P_b , the i_m element of P_m , and SUM .

 Add A to the element of T in the i_b row and the i_m column.

end while

To see that the above algorithm converges and provides a table T with required properties it is sufficient to note that:

1. At each iteration, the row sums of T plus P_b equal the given b marginal.
2. At each iteration, the column sums of T plus P_m equal the given m marginal.
3. At each iteration, an element of P_b or P_m becomes 0.
4. At each iteration, SUM, the sum of the element of P_b , and the sum of the elements of P_m are all equal.

The first two points ensure that the constructed table T will have the specified marginals. The second two points ensure convergence which occurs at a final iteration in which both P_b and P_m have all elements equal to zero and SUM equals zero.

We illustrate the algorithm by tracing through a simple example. We wish to construct a 3×3 table with marginal for b equal to $(.2, .3, .5)$ and marginal for m equal to $(.25, .35, .4)$. For both b and m , we use the rule which chooses the largest value ($r = 1$). The results of each iteration are given in Figure (1). Each iteration of the algorithm is depicted by giving the current value of T in the 3×3 upper left part of the tableau, the current value of P_b in the right margin, the current value of P_m in the bottom margin, and SUM is the single number at the lower right. The values selected by the rules are enclosed in a box.

The “initial state” in Figure (1) shows T set to all zeros, P_b set to the given b marginal, P_m set to the given m marginal, and SUM set to one. The b rule selects the value $v_b = .5$

(with associated index $i_b = 3$, namely the third row). The m rule selects the value $v_m = .4$ (with associated index $i_m = 3$, namely the third column). The minimum of $.4$ and $.5$ is $.4$, and hence $.4$ is inserted in cell $(3,3)$ of T . To get the value of T given at iteration 1, $.4$ is subtracted from the third element of P_b , the third element of P_m , and SUM. Thus, the row marginal 0.5 is updated to $.5-.4=.1$. The column marginal $.4$ is updated to $.4-.4=0$. Note that since the third element of P_m is now zero, our rule will never select the third entry in future iterations so that all entries in the column remain zero save for cell $(3,3)$.

Then the second step comes into action. The largest of the REMAINING row marginals and the largest of the REMAINING column marginals are picked up, and the smallest of the two is inserted in the corresponding cell. The new entry is subtracted from SUM and from the corresponding marginals.

initial state	iteration 1	iteration 2																																																
<table style="width: 100%; border-collapse: collapse;"> <tr><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.20</td></tr> <tr><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.30</td></tr> <tr><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px; border: 1px solid black;">.50</td></tr> <tr style="border-top: 1px solid black;"><td style="padding: 2px 10px;">.25</td><td style="padding: 2px 10px;">.35</td><td style="padding: 2px 10px; border: 1px solid black;">.40</td><td style="padding: 2px 10px;">1.0</td></tr> </table>	.00	.00	.00	.20	.00	.00	.00	.30	.00	.00	.00	.50	.25	.35	.40	1.0	<table style="width: 100%; border-collapse: collapse;"> <tr><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.20</td></tr> <tr><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px; border: 1px solid black;">.30</td></tr> <tr><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.40</td><td style="padding: 2px 10px;">.10</td></tr> <tr style="border-top: 1px solid black;"><td style="padding: 2px 10px;">.25</td><td style="padding: 2px 10px; border: 1px solid black;">.35</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.60</td></tr> </table>	.00	.00	.00	.20	.00	.00	.00	.30	.00	.00	.40	.10	.25	.35	.00	.60	<table style="width: 100%; border-collapse: collapse;"> <tr><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px; border: 1px solid black;">.20</td></tr> <tr><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.30</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td></tr> <tr><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.40</td><td style="padding: 2px 10px;">.10</td></tr> <tr style="border-top: 1px solid black;"><td style="padding: 2px 10px; border: 1px solid black;">.25</td><td style="padding: 2px 10px;">.05</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.30</td></tr> </table>	.00	.00	.00	.20	.00	.30	.00	.00	.00	.00	.40	.10	.25	.05	.00	.30
.00	.00	.00	.20																																															
.00	.00	.00	.30																																															
.00	.00	.00	.50																																															
.25	.35	.40	1.0																																															
.00	.00	.00	.20																																															
.00	.00	.00	.30																																															
.00	.00	.40	.10																																															
.25	.35	.00	.60																																															
.00	.00	.00	.20																																															
.00	.30	.00	.00																																															
.00	.00	.40	.10																																															
.25	.05	.00	.30																																															
iteration 3	iteration 4	iteration 5, final table																																																
<table style="width: 100%; border-collapse: collapse;"> <tr><td style="padding: 2px 10px;">.20</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td></tr> <tr><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.30</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td></tr> <tr><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.40</td><td style="padding: 2px 10px; border: 1px solid black;">.10</td></tr> <tr style="border-top: 1px solid black;"><td style="padding: 2px 10px; border: 1px solid black;">.05</td><td style="padding: 2px 10px;">.05</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.10</td></tr> </table>	.20	.00	.00	.00	.00	.30	.00	.00	.00	.00	.40	.10	.05	.05	.00	.10	<table style="width: 100%; border-collapse: collapse;"> <tr><td style="padding: 2px 10px;">.20</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td></tr> <tr><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.30</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td></tr> <tr><td style="padding: 2px 10px;">.05</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.40</td><td style="padding: 2px 10px; border: 1px solid black;">.05</td></tr> <tr style="border-top: 1px solid black;"><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px; border: 1px solid black;">.05</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.05</td></tr> </table>	.20	.00	.00	.00	.00	.30	.00	.00	.05	.00	.40	.05	.00	.05	.00	.05	<table style="width: 100%; border-collapse: collapse;"> <tr><td style="padding: 2px 10px;">.20</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td></tr> <tr><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.30</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td></tr> <tr><td style="padding: 2px 10px;">.05</td><td style="padding: 2px 10px;">.05</td><td style="padding: 2px 10px;">.40</td><td style="padding: 2px 10px;">.00</td></tr> <tr style="border-top: 1px solid black;"><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td><td style="padding: 2px 10px;">.00</td></tr> </table>	.20	.00	.00	.00	.00	.30	.00	.00	.05	.05	.40	.00	.00	.00	.00	.00
.20	.00	.00	.00																																															
.00	.30	.00	.00																																															
.00	.00	.40	.10																																															
.05	.05	.00	.10																																															
.20	.00	.00	.00																																															
.00	.30	.00	.00																																															
.05	.00	.40	.05																																															
.00	.05	.00	.05																																															
.20	.00	.00	.00																																															
.00	.30	.00	.00																																															
.05	.05	.40	.00																																															
.00	.00	.00	.00																																															

Figure 1: Evident dependence algorithm. Both the column rule and the row rule select the largest element.

Our final table is given at the fifth iteration. The table has the correct marginals and the dependence is, indeed, “evident”. In fact, by Dall’Aglia & Bona (2011) the algorithmic rule just exemplified can be shown to exhibit the strongest dependence between the underlying variables, given the marginals, as it minimizes the entropy of the joint distribution with the underlying marginals.

Figure (2) present a second example. Only the initial state, first iteration, and final table are given. In this example, everything is the same except that the b rule now selects the second largest positive element instead of the largest. To move from initial state to the first

iteration, the value .3 is selected from P_b rather than the value .5. The final table again has the correct marginals and exhibits strong dependence.

initial state				iteration 1				Final Table				
.00	.00	.00	.20	.00	.00	.00	.20	.00	.20	.00	.00	
.00	.00	.00	.30	.00	.00	.30	.0000	.00	.30	.00
.00	.00	.00	.50	.00	.00	.00	.50		.25	.15	.10	.00
.25	.35	.40	1.0	.25	.35	.10	.70		.00	.00	.00	.00

Figure 2: Evident dependence algorithm. Row rule selects the second largest element. Column rule selects the largest.

This second example will not minimize the entropy, but still keeps most of the entries of the joint distribution on the Frechet boundaries hence “a strong dependence” between the underlying variables. All algorithmic rules employed in this paper enjoy the same property of attaining the Frechet boundaries as dictated by the underlying rule. Hence, all distributions created by the algorithm are guaranteed to exhibit dependence.

3 Detailed Description of the Proposed Approach

Section 3.1 presents our basic framework and notation for the data fusion problem. Section 3.2 reviews the GMR approach, casting it as a special case of the general framework. Sections 3.3 and 3.4 detail our general Bayesian model, prior choice, and a simple fitting algorithm. Section 3.5 outlines a more general computational approach.

3.1 Framework for Categorical Data Fusion

In view of the fact that GMR is an important reference paper, we adopt some of the same notation. We label our two target variables b and m so that our goal is inferring the joint distribution of (b, m) . In GMR, b referred to a choice of product to buy and m referred to a chosen media outlet.

Let $D_b = \{(x_i, b_i), i = 1, \dots, N_b\}$, denote a data set of observations on the target variable b and the common variables x . $D_m = \{(x_i, m_i), i = 1, \dots, N_m\}$ denotes the data set of observations on the other target variable m and x . $D_{b,m} = \{(x_i, b_i, m_i), i = 1, \dots, N_{b,m}\}$ is a

data set of fully fused data recording all three of b , m and x . Throughout, we are motivated by the observation that it is often the case in practice that D_b and D_m are “large”, while $D_{b,m}$ is “small”.

We view data fusion as a means to form inferences about the joint distribution of (b, m) using the information in the entire data $D = (D_b, D_m, D_{b,m})$. Estimates of the joint distribution of (b, m) can then be used to solve whatever decision problem is required by the application. For example, in media planning, media choices that have a high proportion of viewers who purchase the advertised product category may be considered desirable.

If we were to assume independence of b and m conditional on x , the marginal joint distribution of (b, m) would be

$$p(b, m) = \int p(b, m | x) p(x) dx = \int p(b|x) p(m|x) p(x) dx. \quad (1)$$

The basic idea of this paper is to introduce a set of parameters, α , that allow us to construct a rich set of possible joint distributions from the conditional marginals $p(b | x)$ and $p(m | x)$. We define $T(p(b | x), p(m | x), \alpha)$ to be a joint distribution for (b, m) such that, for all α , the marginal of b is $p(b | x)$ and the marginal of m is $p(m | x)$. We then let

$$p(b, m | x, \alpha) = T(p(b | x), p(m | x), \alpha) \equiv T(x, \alpha). \quad (2)$$

With this definition we have,

$$p(b, m | \alpha) = \int T(x, \alpha) p(x) dx \quad (3)$$

As in Section 2, the notation “ T ” is meant to suggest the common two-way table representation of the joint distribution of (b, m) . We will use the same symbol T to denote a joint represented in a manner appropriate for the expression in equation (3) and the corresponding table whose i, j element is the probability that b takes on its i^{th} possible value and m takes on its j^{th} possible value. We abuse notation slightly by writing $p(b, m | x, \alpha) = T(x, \alpha)$ when T is a table, but clearly the information in the left and right hand sides is the same. If we want to evaluate the left hand side at a particular (b, m) pair, we select the corresponding entry from the table T . In the above equations we have used the “ dx ” notation appropriate for Lebesgue measure. The obvious adjustments may be made in the (quite common) case where some components of x are categorical.

There are many ways to incorporate conditional dependence by replacing (1) with some model for the conditional joint distribution $p(b, m | x)$. For example, Rassler (2002) introduces

a prior distribution that captures some view of dependence for the case of multivariate normal (b, m, x) . Since our goal is the fusion of categorical variables, the multivariate normal distribution is inappropriate.

3.2 Relaxing the Conditional Independence Assumption – The GMR Approach

At this point it is useful to briefly review what GMR proposed for binary variables in terms of the structure given above.

In the binary case we can assume both b and m have possible outcomes given by the set $\{1, 2\}$. The marginals for b and m are then captured by $p_b = Pr(b = 2)$ and $p_m = Pr(m = 2)$. Under independence, the joint distribution of (b, m) is represented by the 2×2 table $P = \{p_{ij}\} = \{Pr(b = i, m = j)\}$ ($i, j \in \{1, 2\}$).

$$P = \begin{bmatrix} (1 - p_b)(1 - p_m) & (1 - p_b)p_m \\ p_b(1 - p_m) & p_b p_m \end{bmatrix}. \quad (4)$$

A departure from independence is suggested in GMR by introducing a parameter λ ($|\lambda| \leq 1$). Let,

$$a = \begin{cases} \lambda \min((1 - p_b)p_m, p_b(1 - p_m)), & \lambda > 0 \\ \lambda \min((1 - p_b)(1 - p_m), p_m p_b), & \lambda < 0. \end{cases}$$

The scalar a can be used to perturb the array P given above in (4) to represent a new multinomial distribution with dependence:

$$T(p_b, p_m, \lambda) = \begin{bmatrix} (1 - p_b)(1 - p_m) + a & (1 - p_b)p_m - a \\ p_b(1 - p_m) - a & p_b p_m + a \end{bmatrix}. \quad (5)$$

If $|\lambda| < 1$, then $T(p_b, p_m, \lambda)$ will constitute a valid multinomial distribution. Positive values of λ will provide for positive conditional dependence and vice versa. The parametrization in (5) will preserve the marginals of b and m while accommodating a specific degree of conditional dependence indexed by λ .

If our marginals for b and m are now conditional on x with $p_b(x) = Pr(b = 2 | x)$ and $p_m(x) = Pr(m = 2 | x)$, then we obtain the GMR approach by letting

$$p(b, m | x, \lambda) = T(p_b(x), p_m(x), \lambda).$$

This has the some structure as equation (2) of Section 3.1 with the GMR parameter λ corresponding to the parameter α and $p_b(x)$ and $p_m(x)$ representing $p(b | x)$ and $p(m | x)$ respectively.

The GMR prior on λ is

$$p(\lambda) \propto \frac{1}{(1 + |\lambda|)^\beta}$$

This prior concentrates on values close to zero, which corresponds to conditional independence. GMR demonstrated the superiority of their approach over competing methods for fusion of binary variables. In the appendix we show that the GMR model is a special case of the more general approach developed in this paper.

In the 2×2 case, given the marginals, taking control of one cell in the table determines all the other cells. When dimensions of the table are higher than 2×2 , GMR's method does not extend in an obvious way. We propose an alternative procedure that can accommodate tables of any dimension.

3.3 A New Fusion Approach for Multi Level Multinomials: Evident Dependence Mixture

In the notation of equation (1), we introduce a model $p(b, m | x, \alpha)$ which works for categorical (b, m) . We use the large data sets D_b and D_m to estimate $p(b | x)$ and $p(m | x)$. We then use the smaller data set $D_{b,m}$ to estimate α . Because $D_{b,m}$ is assumed small, we augment this information with a prior on α whose choice is straightforwardly motivated by the goal of shrinking towards the conditional independence solution.

As discussed in Section 2, a choice of row rule and column rule gives us an evident dependent table given marginals for b and m . Let k index a set of choices for both the row and column rules, $k = 1, 2, \dots, K$. For example, in Section 4 below, we consider five possible row rules and five possible column rules giving $k = 25$ possible rule combinations. Let $T_k(p_b, p_m)$ denote the table resulting from a b marginal p_b , a m marginal p_m , and rule choice k .

Let $T_o(p_b, p_m)$ be the table constructed by letting b and m be independent. Our overall table is a mixture of the evident dependence tables and the independence table:

$$T(p_b, p_m, \alpha) = \sum_{k=0}^K \alpha_k T_k(p_b, p_m) \tag{6}$$

where $\sum_{k=0}^K \alpha_k = 1$ and $\alpha_k \geq 0$. Clearly, since each T_k has marginals p_b and p_m , so does $T(p_b, p_m, \alpha)$. Marginal preserving is strongly advocated by Kiesel & Rassler (2007).

We now consider construction of our table conditional on x . The conditional marginal models $p(b|x)$ and $p(m|x)$, may be represented as vectors $p_b(x)$ and $p_m(x)$ giving conditional probabilities for each of the possible outcomes as in the unconditional case above. We then have

$$p(b, m | x, \alpha) = T(x, \alpha) = T(p_b(x), p_m(x), \alpha). \quad (7)$$

The first “equality” in the expression above is simply the notation of Section 3.1. The second equality says we obtain our conditional model for the joint distribution of (b, m) by plugging the conditional marginals into the construction of equation (6).

Our overall estimation scheme for fusing b and m consists of the following steps:

1. Estimate $p(b|x)$ and $p(m|x)$ using D_b and D_m .
2. Plug in the models obtained from the previous step to obtain $p(b, m | x, \alpha) = T(x, \alpha)$ using equations (6) and (7) above.
3. Let $p(\alpha)$ denote the prior density of α .
Estimate α using $D_{b,m}$ and $p(\alpha | D_{b,m}) \propto p(\alpha) \prod_{i=1}^{N_{b,m}} p(b_i, m_i | x_i, \alpha)$.
We let the estimate of α , $\hat{\alpha}$, be the posterior mode.
4. Estimate the table corresponding to $p(b, m)$ by $\hat{T} = \frac{1}{N} \sum_{i=1}^N T(x_i, \hat{\alpha})$.

The problem of estimating the marginal models in step 1 is a very common one which can be handled using standard statistical methodology (e.g. multinomial logit). The x 's used in fourth step are from all three data sets D_b , D_m , and $D_{b,m}$ so that typically $N = N_b + N_m + N_{b,m}$. We average over x as an approximation to the integral $\int T(x, \alpha) p(x) dx$ in equation (3).

This construction is a simple way to achieve our goal of combining the information from all three data sets. We could compute the likelihood from all three data sets using the model implied by our mixture, but given our assumptions about the size of the data sets, we feel that plugging in the estimated marginal models in the second step is a reasonable approach which greatly simplifies the methodology. We plug in the posterior mode in step 4 to obtain a point estimate, but in this case we are ignoring uncertainty under that assumption that $D_{b,m}$ is small. More detailed discussion of the structure of the likelihood and inference options is given in Section 3.5.

The choice of prior $p(\alpha)$, will matter. A reasonable choice for the form of $p(\alpha)$ is the standard Dirichlet prior. The choice of prior is motivated by the belief that conditional independence may be a reasonable assumption and simplicity. In our examples, we use the prior $p(\alpha) \propto \alpha_0^2$. This prior favors α vectors with α_0 , the weight on the conditional independence table, close to one. We “regularize” our approach by pushing our solution towards simple conditional independence. Further dependence ($\alpha_0 < 1$) is introduced to the fit only if suggested by the information in $D_{b,m}$. In the spirit of the Lasso (Tibshirani 1996), we typically find that at the mode many elements of α are set to zero. Of course, the appropriate amount of regularization is always an issue in practice.

3.4 Prior and Modeling Choices

In our examples, we use the multinomial logit for the families $p(b|x, \theta_b)$ and $p(m|x, \theta_m)$. The variables in x are categorical and represented in the usual way as sets of dummy variables. Using dummies in this way makes the multinomial logit fairly flexible. Given large data sets, it may be of interest to explore the use of more flexible modeling approaches for the estimation of the marginal models.

A nice feature of our approach is the clear motivation for a prior which places weight on large values of α_0 , the mixture probability of the conditional independence table. In Section 3.3 we suggested the prior $p(\alpha) \propto \alpha_0^2$. This choice “regularizes” our approach by pushing our inference towards the simple conditional independence model which often works very well in practice. The appropriate degree of shrinkage is always a question to be considered. If we use priors of the form $p(\alpha) \propto \alpha_0^\gamma$, what is a good choice of γ ?

There is also the number of column and row rules to be considered. These choices, in turn, determine the dimension of α . Given the nature of our evident-dependence algorithm, it may also make sense to put more weight on α_i which are weights for tables constructed using using low values for the column and row rules. This is because these rules put in the most dependence. A search for a parsimonious model which captures dependence could reasonably focus on such tables.

An alternative approach to fusion would involve the use of latent continuous variables to model the categorical variables. This construction has been very popular in the Bayesian literature. The use of such latent variables would have the advantage that standard dependence models for continuous variables could be used. Typically such models are computed using an MCMC algorithm in which the latent variables are draw as part of larger Gibbs sampler.

However, drawing these latents may induce substantial autocorrelation in the MCMC chain. Also, it may be difficult to express prior beliefs in terms of latent variables. We feel the simplicity of our approach with natural prior choices and marginal preservation makes it an appealing alternative.

3.5 Approximate and Exact Inference

Implementation of our approach involves choices of the marginal models $p(b|x)$ and $p(m|x)$. These marginal models will typically have associated parameters which we denote by θ_b and θ_m , giving $p(b|x, \theta_b)$ and $p(m|x, \theta_m)$.

We have,

$$p(b, m | x, \theta_b, \theta_m, \alpha) = T(p(b | x, \theta_b), p(m | x, \theta_m), \alpha),$$

which is equation (7) with the marginal model parameters explicitly indicated.

Because the evident dependence algorithm preserves marginals we have

$$p(b | x, \theta_b, \theta_m, \alpha) = p(b | x, \theta_b), \quad p(m | x, \theta_b, \theta_m, \alpha) = p(m | x, \theta_m).$$

Our likelihood is then,

$$\begin{aligned} L(\theta_b, \theta_m, \alpha | D_b, D_m, D_{b,m}) &= \left[\prod_{i=1}^{N_b} p(b_i | x_i, \theta_b, \theta_m, \alpha) \right] \left[\prod_{i=1}^{N_m} p(m_i | x_i, \theta_b, \theta_m, \alpha) \right] \left[\prod_{i=1}^{N_{b,m}} p(b_i, m_i | x_i, \theta_b, \theta_m, \alpha) \right] \\ &= \left[\prod_{i=1}^{N_b} p(b_i | x_i, \theta_b) \right] \left[\prod_{i=1}^{N_m} p(m_i | x_i, \theta_m) \right] \left[\prod_{i=1}^{N_{b,m}} p(b_i, m_i | x_i, \theta_b, \theta_m, \alpha) \right] \\ &\equiv L_b(\theta_b) L_m(\theta_m) L_{b,m}(\theta_b, \theta_m, \alpha). \end{aligned}$$

where the three terms correspond to the three sets D_b , D_m and $D_{b,m}$. We see that the margin preserving property gives the likelihood an appealing factorization.

In cases where N_m and N_b are large, we clearly expect inference for θ_b and θ_m to be dominated by the likelihood terms L_b and L_m respectively. Since these likelihoods are simply those of the marginal models, we have a strong motivation for the plug-in approach of Section 3.3.

If we wish more exact inference, the obvious Gibbs sampler consists of the draws:

$$\theta_b | \theta_m, \alpha, D, \quad \theta_m | \theta_b, \alpha, D, \quad \alpha | \theta_b, \theta_m, D$$

where D denotes all of the data. The draws for θ_b and θ_m can be done using Metropolis-within-Gibbs steps where we generate proposals based on the marginal model likelihood and

prior and then accept using the full likelihood and prior. For commonly used models such as the multinomial logit, simple normal based approximations to the marginal model posterior generate good proposals.

In drawing α we need to be aware that the mode may have some elements set to zero. This is due to the highly non-linear nature of the evident dependence algorithm. As mentioned above, we view this as a good thing as it naturally pushes us towards parsimonious solutions.

4 Empirical Results

In this section we apply our evident dependence fusion (ED) method to survey data. In Section 4.1 we describe the data.

We start with a large fully-fused data set. From the many variables in our data set, we choose variables to play the roles of b , m , and x . We subsample our data into three subsets and remove b from one set to create D_m , remove m from another set to create D_b , and leave all of b , m , and x in the third to create $D_{b,m}$. Given the three sets, we apply ED to obtain an estimate of the table giving the joint distribution of (b, m) . We then compare this estimate to the “true” table where we use the sample table using all our data as the true table. The Kullback-Leibler divergence is used to measure the difference between the true table and the estimated table. Section (4.2) reviews the Kullback-Leibler divergence and provides a simple way to interpret divergence values.

In Section 4.3 we repeatedly subsample our data to draw the sets D_b , D_b , $D_{b,m}$. For each draw, we compare the ED estimate with the true table and the estimate obtained using the assumption of conditional independence (CI). CI fusion corresponds to ED with α_0 set to one. We report results for two choices of the (b, m) pair. The first pair is chosen because we expect the pair to exhibit strong dependence while the second pair is chosen because we expect weak dependence.

We find that, with a small amount of fused data, the ED method dramatically improves our estimation for the pair where there is strong dependence and does no harm in the pair with weak dependence.

4.1 The Data

Our data comes from a survey of British consumers conducted in 1998 by the British Market Research Bureau. We have observations from 24,497 households. Broadly speaking, there are four types of variables: demographics, attitudinal, product or product category usage, and media exposure. All of our variables are categorical.

The data set has nineteen demographic variables. We used only ten of the nineteen because the data provider BMRB advised us that these were the variables they most commonly use in their data fusion practice. Variables in x are:

1. income, 12 levels
2. social grade, 6 levels
3. education, 10 levels
4. age, 7 levels
5. working status, 5 levels
6. marital status, 5 levels
7. number of children in household, 6 levels
8. number of adults in household, 6 levels
9. gender of household respondent, 2 levels
10. geographic region of household, 11 levels

For a (b, m) pair which we expect to be highly dependent, we chose b to measure the amount of cigarette smoking and m to measure the degree to which the respond agrees with the statement “Smoking should be banned in public places”.

The “cig” b has the six possible responses: Not Stated, 40 A Day Or More, More Than 20 But Less Than 40 A Day, More Than 10 But Less Than 20 A Day, More Than 5 But Less Than 10 A Day, Under 5 A Day.

The “ban” m has the seven possible responses: Definitely Agree, Tend To Agree, Neither Agree Nor Disagree, Tend To Disagree, Definitely Disagree, Not Applicable, Not Stated.

For a pair where we expect low dependence, we again used the the “amount of smoking” for b . We chose m to the variable measuring the respondent’s viewing of the television program “World in Action”. Possible responses are: Specially Chose To Watch, Watch Because Someone In, Watch When Nothing Better, I Don’t Watch It, Not Stated.

4.2 The Kullback-Leibler Divergence

To gauge the success of a fusion method, we need to quantify how different the true table is from the estimated table. We shall use the Kullback-Leibler divergence as a distance measure. Let n_b and n_m denote the number of levels of b and m respectively. Then the number of possible outcomes is $I = n_b \times n_m$. Let p_i be the true probability of the i^{th} outcome and \hat{p}_i be the estimate. Then the Kullback-Leibler divergence (KL) is given by

$$\text{KL}(p, \hat{p}) = \sum_{i=1}^I p_i \log(p_i/\hat{p}_i)$$

To give the reader a sense of what a KL value means, we consider a binary b and a binary m whose joint distribution is then represented by the standard 2×2 table. We compute the KL divergence between the two tables

$$T = \begin{bmatrix} .25 & .25 \\ .25 & .25 \end{bmatrix} \quad \text{and} \quad T(\epsilon) = \begin{bmatrix} .25 + \epsilon & .25 - \epsilon \\ .25 - \epsilon & .25 + \epsilon \end{bmatrix}$$

where $\epsilon \in (0, .25)$. When ϵ is zero, b and m are independent and $T(\epsilon) = T$. As ϵ increases, the dependence in $T(\epsilon)$ increases and $T(\epsilon)$ moves away from T .

Figure (3) plots ϵ vs. $\text{KL}(T, T(\epsilon))$. The two panels differ only in the scale on the axes. ϵ values of .01, .02, .03, .04, and .05 correspond to KL values of .0008, .0032, .0072, .013, and .02. KL values of .005, .01, and .02 correspond to ϵ values of .025, .035, and .05. Hence, the convexity of $\text{KL}(T, T(\epsilon))$ means that a small change in ϵ may result (or be implied from) relatively large changes in KL. In other words, a significant (qualitative) change from total independence to strong dependence will express itself in a relatively large change in the KL divergence measure.

4.3 Evident Dependence Fusion: A Sampling Experiment

In this section we report on a sampling experiment designed to compare the estimates obtained from ED with the true table.

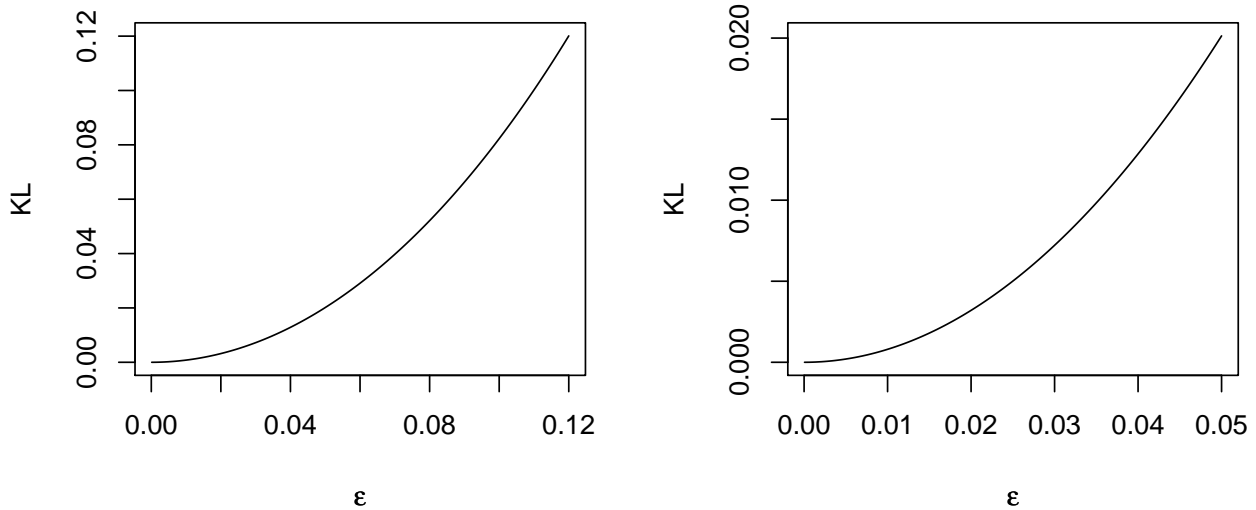


Figure 3: Plot of ϵ vs. $KL(T, T(\epsilon))$. A KL value of .1 could be considered large because it corresponds to an ϵ value of about .1.

Each draw of our sampling experiment randomly partitions the observations into our three sets D_b , D_m , and $D_{b,m}$. We allocate 1% of the observations to the fused set $D_{b,m}$ and the remaining observations are equally split between D_b and D_m so that each gets 49.5%. Our complete data set has 24,497 observations. Thus, in each draw, $D_{b,m}$ has 245 observations and D_b and D_m both have 12,126 observations. We did 1,000 draws. We did the experiment twice. The first time we used just the first four of our conditioning variables (income, social grade, education, age) in x and the second time we used all ten.

We used the multinomial logit model to estimate the conditional marginal models $p(b|x)$ and $p(m|x)$. Our categorical x variables were expressed as dummies in the usual way. For both b and m , the set of rules $\{1, 2, 3, 4, 5\}$ were considered so that the total number of evident dependence tables J is equal to 25. Thus the total number of tables in our mixture (equation (6)) is 26.

The results using four variables in x are given in Figure (4). Each boxplot depicts 1,000 Kullback-Leibler divergences (distances) between the true table and the estimated table. The 1,000 distances correspond to the 1,000 draws of the three data sets in our sampling experiment. The first boxplot reports the results for the pair (cig,ban) using the estimation method which takes just the 245 observations of fused data in the draw of $D_{b,m}$ and computes

the sample two-way table. The next two boxplots report results for the (b, m) pair (cig, ban), with the first of the pair giving CI fusion results and the second giving ED fusion results. The last pair of boxplots report results of the CI and ED fusion for the pair (cig, world).

The first three boxplots report results for the pair (cig,ban). We see that while conditional independence fusion reduces the sampling variability, the overall level of the fusion is not improved beyond simply taking the sample estimate using the small amount of fused data. The third boxplot shows that evident dependence fusion dramatically improve the quality of the estimation. For every sample the KL divergence obtained using evident dependence is smaller than that obtained using conditional independence. The mean KL for conditional independence fusion is .11 while the mean KL for evident dependence fusion is .037 which is $(1/3)$ as large. A KL of .11 corresponds to an ϵ of .12 and a KL of .037 corresponds to an ϵ of .05. Clearly, on both the KL scale and the ϵ scale we see that the evident dependence fusion is dramatically better. The last pair of boxplots (the results for the pair (cig, world)) show that using the richer evident dependence fusion model does not degrade the results when conditional independence is already successful.

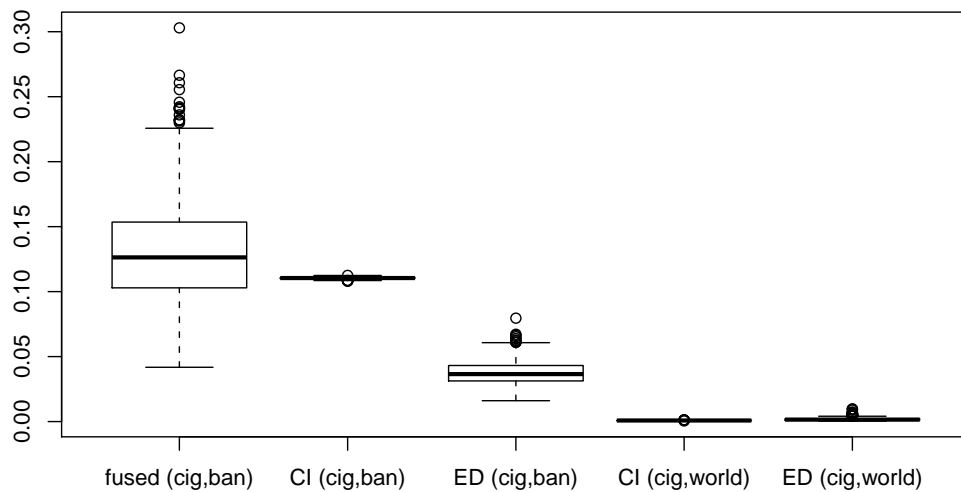


Figure 4: Sampling results for two pairs of interest, four conditioning variables.

Figure (5) reports the results obtained using all 10 variables in x . The format is the same as in Figure (4) except that the first boxplot is deleted. The results for CI fusion are virtually

identical to those we obtained with just four x variables. Again, the ED fusion is dramatically better than CI fusion for the (cig,ban) pair and very similar for the (cig, world) pair.

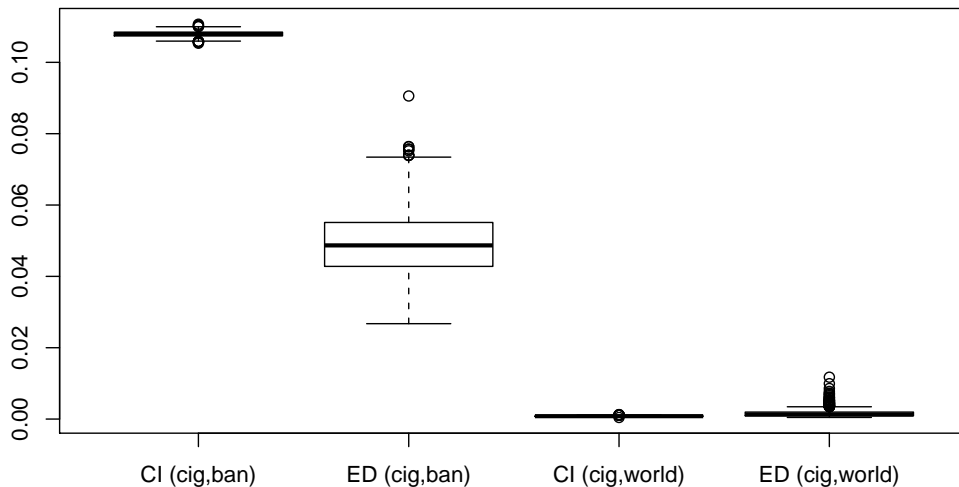


Figure 5: Sampling results for two pairs of interest, ten conditioning variables.

In Figure (6) we report the posterior mode estimates of the mixture weight for the conditional independence table (α_0 in Section 3.3) in the evident dependence fusion. The boxplot on the left gives the 1,000 estimates of α_0 obtained from the 1,000 sub-samples for the (cig, ban) pair. The boxplot on the right gives the weight estimates for the (cig, world) pair. As one would expect, the data directs us to give more weight to conditional independence in the (cig, world) pair.

While KL is a useful summary of the fusion results, we can also look at quantities of interest one might compute from the estimated table and see how they differ as the table estimates vary. There are many possible quantities we could look at. We choose to look at $\text{Prob}(m \text{ is either "Definitely Agree" or "Tend to Agree" } | \text{ levels of } b)$. We see how the probability that the respondent basically agrees that smoking should be banned in public places depends on their response to the query about how much they smoke.

Figure (7) compares the ED estimate of $\text{Prob}(m \text{ is either "Definitely Agree" or "Tend to Agree" } | \text{ levels of } b)$ with those obtained using CI and the true value. The horizontal axis is labeled with b_1, b_1, \dots, b_6 , denoting the six possible levels of b . The levels of b are

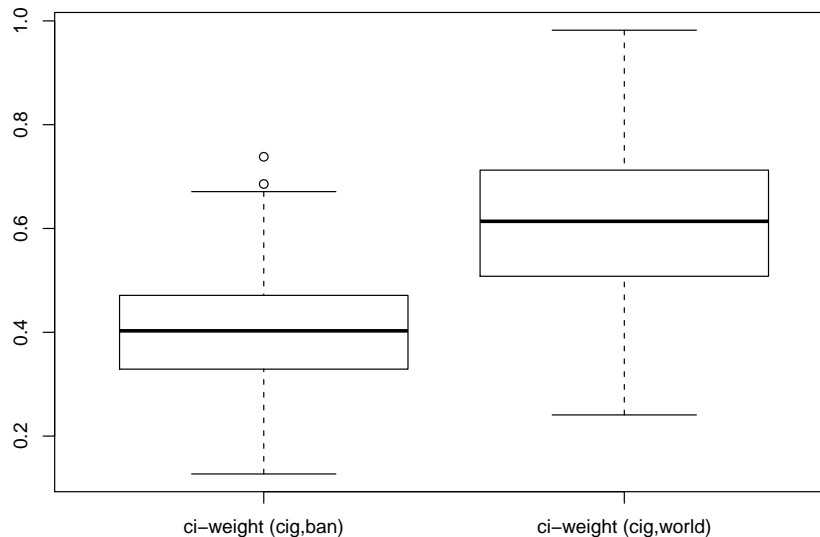


Figure 6: Posterior mode estimates of mixture weight.

given in Section 4.1. **b1** corresponds to the response “Not Stated” while **b2-b6** represent decreasing amounts of smoking, going from “40 a day or more” down to “Under 5 a day”.

Above each b level, there are two boxplots and a dot. The dot gives the conditional probability computed from the true (b, m) table. The boxplot on the left gives the 1,000 conditional probabilities computed from the estimated tables using CI fusion for the 1,000 sub-samples. The boxplot on the right gives the results from the ED fusion. Clearly, the boxplots on the right are much closer to the truth (the dot). From the dots, we see that the truth is that a b of “Not Stated” makes it much more likely that a person will support a ban. This strong dependence is clearly captured in the evident dependence fusion, but almost entirely lost by the conditional dependence fusion.

5 Conclusion

In this paper we discussed a new fusion methodology that can extract much more information from the marginal distributions of the target variables b and m than the common conditional independence fusion (CI). Our methodology requires a small sample of fused data, and uses

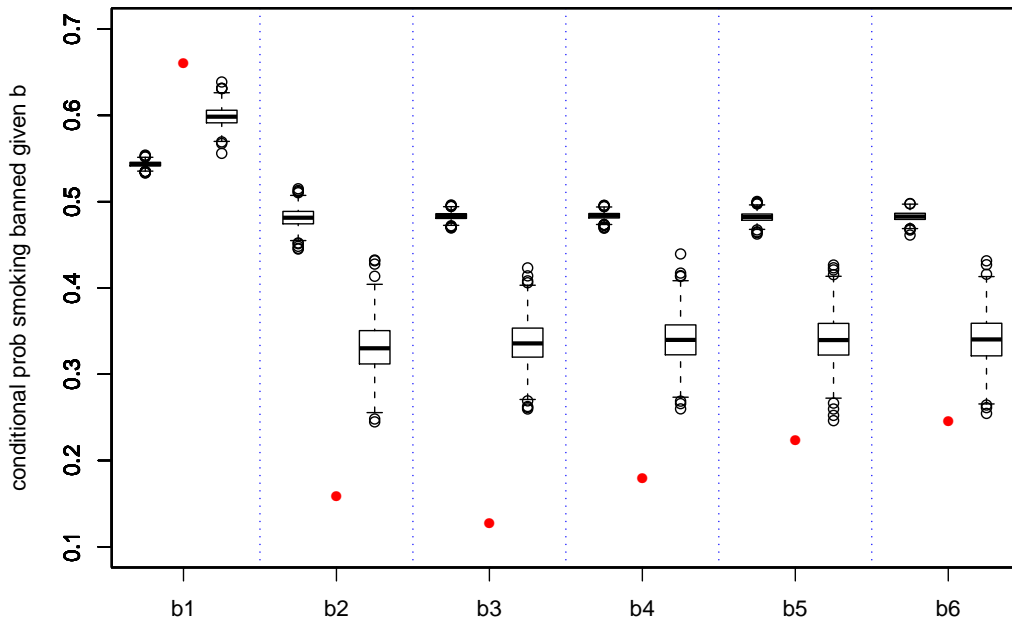


Figure 7: Sampling experiment results for ban given cig. Levels of b on the horizontal axis. Estimation of $\text{Prob}(m \text{ is either "Definitely Agree" or "Tend to Agree" } | \text{ levels of } b)$ on the vertical axis.

it to estimate the joint using a Bayesian approach to mixture modeling. In case there is no such sample, or the sampled fused data dictate it, our methodology simply boils down to CI fusion. In addition, it was shown that the approach to fusion taken in this paper generalizes GMRs earlier reported fusion technique based on partially fused data that applies to 2x2 tables. Hence, the methodology advocated in this paper enjoys a wide generality and a high degree of accuracy. As discussed in GMR, in theory, if the number of common variables x is VERY large, then CI fusion should work quite accurately. The higher the dimension of the space of x the more likely it is that the target variables b and m are indeed conditionally independent given x . However, in reality, a great deal of the companies practicing data fusion use “constrained” fusion, namely, using ON PURPOSE, for practical reasons, a relatively few common variables x . Thus, the Evident Dependence fusion technique should also prove quite practical. The authors will be happy to provide readers with code and a user guide upon request.

Note that in fact much more information can still be extracted from the (b, m) marginals if we allow considering lower degrees of dependence between the target variable than in the Evident Dependence case. This requires a non-trivial set of definitions and complicated algorithms to cover the entire spectrum from full independence to the strongest dependence given the marginals. This challenging task is left for future research. It is important to note that the main problem of estimating the joint distribution from its marginals is addressed in other areas. There the problems are similar but NOT IDENTICAL to data fusion addressed in this paper. In the bio–statistical context, related problems have been addressed over 30 years ago (see, e.g. Chen & Feinberg (1974)) and Chen & Feinberg (1976)). Other areas where very important and useful methodologies on estimating joint distributions given the marginals (and some added information) have been developed are confidentiality and disclosure and ecological inference. In confidentiality and disclosure we note Feinberg & Makov (1998), Feinberg, Makov & Steele (1998), Feinberg, Makov, Meyer & Steele (2001), Slavkovic & Feinberg (2005), and Feinberg (2006). Ecological inference deals with the challenging problem of whether relationships observed for groups necessarily hold for individuals. The methodological problem there is similar (but not identical) to data fusion, where joint distributions are sought between two discrete variables where the available information is sample marginals on one variable and population marginals on the other variable. The important paper by Wakefield (2004) and the book by King, Rosen, and Tanner (2004) are excellent sources of the methodologies existing in this field.

Finally, we would like to briefly mention the great potential of copulas to data fusion. Elidan (2010), uses copulas (e.g. Nelsen (2007)) to provide a general framework of constructing

multivariate distributions given their marginals. This framework is based on the well-known Sklar's Theorem (Sklar (1959)) that any multivariate distribution can be represented as a copula function of its marginals. Copulas have diversified and far reaching applications as is evident for instance from Embrechts P. & A. (2003) and from Accioly & Chiyoshi (2004). The copula approach accommodates well the modeling of dependence patterns between the marginal variables. Copulas are mostly employed for continuous distributions where copulas are unique. Uniqueness of copulas generally does not hold for discrete distributions like the multinomial distribution addressed in this paper. Clearly, the approach of this paper may be viewed as a kind of copula in that we stitch together marginals to form a joint. We believe data fusion can benefit a lot from relevant research into copulas for (constrained) discrete distribution. Although out of the scope of this paper, we intend to investigate the relevance and merits of the methodology developed in this paper to copulas of multinomials.

6 Appendix: The GMR Approach as a Special Case of Evident Dependence

We now show that the GMR approach as a special case of Evident Dependence

As discussed in Section 3.2, the GMR table is

$$\begin{bmatrix} (1 - p_b)(1 - p_m) + a & (1 - p_b)p_m - a \\ p_b(1 - p_m) - a & p_b p_m + a \end{bmatrix}$$

Consider the case where $(1 - p_b)p_m < (1 - p_m)p_b$. This is equivalent to $p_m < p_b$. In this case, the (2, 2) element of GMR table is $p_b p_b + \lambda(1 - p_b)p_m = \lambda p_m + (1 - \lambda)p_b p_m$. Clearly by appropriate choice of B rule and M rule, we can construct the evident dependence table

$$\begin{bmatrix} (1 - p_b) & 0 \\ p_b p_m & p_m \end{bmatrix}.$$

Hence, if we consider an α that puts weight $1 - \omega$ on the conditional independence table and ω on the evident dependence table given above, the (2, 2) element is $(1 - \omega)p_b p_m + \omega p_m$.

Thus, the two approaches give the same set of possible (2, 2) elements. Since both marginals are fixed, the entire table is determined by any one element so we see that the two approaches give the same set of possible tables.

If λ is positive and $p_b > p_m$ we proceed similarly except that we use the evident dependence table with a zero in (2, 1) position. For negative λ , the GMR approach depends on whether $(1 - p_b)(1 - p_m) < p_b p_m$. In a manner analogous to the above the generated tables correspond to mixing the conditional independence table with evident dependence tables have zeros in either the (1, 1) or (2, 2) position.

References

- Accioly, R. & Chiyoshi, F. (2004), ‘Modeling dependence with copulas: a useful tool for field development decision process’, *Journal of Petroleum Science and Engineering* **44**, 83–91.
- Chen, T. & Feinberg, S. (1974), ‘Two-dimensional contingency tables with both completely and partially cross-classified data’, *Biometrics* **30**, 629–642.
- Chen, T. & Feinberg, S. (1976), ‘The analysis of contingency tables with incompletely classified data’, *Biometrics* pp. 629–642.
- Dall’Aglia, G. & Bona, E. (2011), ‘The minimum of the entropy of a two-dimensional distribution with given marginals’, *Electronic Journal of Statistics*.
- Dobra, A. & Feinberg, S. (2001), ‘Bounds for cell entries in contingency tables induced by fixed marginal totals with applications to disclosure limitation’, *Statistical Journal of the United Nations ECE* **18**, 363–371.
- Dobra, A. & Feinberg, S. (2003), Bounding entries in multi-way contingency tables given a set of marginal totals, in Y. Haitovsky, H. Lerche & Y. Ritov, eds, ‘Foundations of Statistical Inference: Proceedings of the Shores Conference 2000’, Phisica-Verlag, pp. 3–16.
- Elidan, G. (2010), Copula bayesian networks, in J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel & A. Culotta, eds, ‘Advances in Neural Information Processing Systems 23’, MIT Press, pp. 559–567.
- Embrechts P., L. F. & A., M. (2003), Modeling dependence with copulas and applications to risk management, in ‘Handbook of Heavy Tailed Distributions in Finance’.
- Feinberg, S. (2006), ‘Privacy and confidentiality in an e-commerce world: Data mining. data warehousing, matching and disclosure limitation’, *Statistical Science* **21**(2), 143–154.
- Feinberg, S. & Makov, U. (1998), ‘Confidentiality, uniqueness and disclosure limitation for categorical data’, *Journal of Official Statistics* **14**, 385–397.

- Feinberg, S., Makov, U., Meyer, M. & Steele, R. (2001), Computing the exact distribution for a multi-way contingency table conditional on its marginal totals, *in* S. A.K.M.E, ed., ‘Data Analysis from Statistical Foundations: A Festschrift in Honor of the 75th Birthday of D. A. S. Fraser’, Nova Science Publishers, pp. 145–165.
- Feinberg, S., Makov, U. & Steele, R. (1998), ‘Disclosure limitation using perturbation and related methods for categorical data’, *Journal of Official Statistics* **14**, 485–511. (With discussion by P. Kooiman and a response.).
- Gilula, Z., McCulloch, R. & Rossi, P. (2006), ‘Direct data fusion’, *Journal of Marketing Research* pp. 1–22.
- Kadane, J. (1978), Some statistical problems in matching data files, *in* ‘1978 Compendium of Tax Research’, Office of the Treasury, pp. 159–171.
- Kamakura, W. & Wedel, M. (1997), ‘Statistical data fusion for cross-tabulation’, *Journal of Marketing Research* **34**, 485–498.
- Kamakura, W. & Wedel, M. (2000), ‘Factor analysis and missing data’, *Journal of Marketing Research* **37**, 490–498.
- Kiesel, H. & Rassler, S. (2007), ‘How valid can data fusion be’, *Journal of Official Statistics*.
- Nelsen, R. (2007), *An Introduction to Copulas*, Springer.
- Rassler, S. (2002), *Statistical Matching*, New York: Springer.
- Rogers, W. (1984), ‘An evaluation of statistical matching’, *Journal of Business and Economic Statistics* **2**, 91–102.
- Rubin, D. (1986), ‘Statistical matching using file concatenation with adjusted weights and multiple imputations’, *Journal of Business and Economic Statistics* **4**, 87–94.
- Sklar, A. (1959), ‘Fonctions de repartition a n dimensions et leurs marges’, *Publication de l’Institut de Statistique de l’Université de Paris* **8**, 229–231.
- Slavkovic, A. & Feinberg, S. (2005), ‘Preserving the confidentiality of categorical statistical data bases when releasing information for association rules’, *Data Mining and Knowledge Discovery* **11**, 155–180.
- Tchen, A. (1980), ‘Inequalities for distributions with given marginals’, *Annals of Probability* **8**, 814–827.