BAYESIAN DISCLOSURE RISK ASSESSMENT: PREDICTING SMALL FREQUENCIES IN CONTINGENCY TABLES

JONATHAN J. FORSTER, EMILY L. WEBB

ABSTRACT

We propose an approach for assessing the risk of individual identification in the release of categorical data. This requires the accurate calculation of predictive probabilities for those cells in a contingency table which have small sample frequencies, making the problem somewhat different from usual contingency table estimation, where interest is generally focussed on regions of high probability. Our approach is Bayesian and provides posterior predictive probabilities of identification risk. By incorporating model uncertainty into our analysis, we can provide more realistic estimates of disclosure risk for individual cell counts than are provided by methods which ignore the multivariate structure of the data set.

Southampton Statistical Sciences Research Institute
Methodology Working Paper M07/05

Bayesian Disclosure Risk Assessment: Predicting Small Frequencies in Contingency Tables

Jonathan J Forster†

University of Southampton, Southampton, UK.

Emily L Webb

The Institute of Cancer Research, London, UK

Summary. We propose an approach for assessing the risk of individual identification in the release of categorical data. This requires the accurate calculation of predictive probabilities for those cells in a contingency table which have small sample frequencies, making the problem somewhat different from usual contingency table estimation, where interest is generally focussed on regions of high probability. Our approach is Bayesian and provides posterior predictive probabilities of identification risk. By incorporating model uncertainty into our analysis, we can provide more realistic estimates of disclosure risk for individual cell counts than are provided by methods which ignore the multivariate structure of the data set.

Keywords: Categorical data; Identification; Model uncertainty; Prediction

1. Introduction

Suppose that an agency releases data on a sample of individuals from a population, and that the sample data consists of the values of a number of categorical variables, recorded for each individual in the sample. Then, the sample data can be expressed as a a multiway contingency table. One important form of identification risk occurs when there are sample cell counts of 1 (uniques) in the marginal table representing the cross-classification of individuals by a subset of v *key* variables (those variables whose values in the population are available to a potential intruder from a source external to the released data under consideration). If the intruder can determine, with confidence, that a sample unique in the contingency table of key variables, is also unique in the population, then this individual can be identified and the data release allows disclosure of the values of the remaining (non-key) variables for this individual. See Bethlehem et al (1990) for more detailed discussion.

Common measures of disclosure risk are based on probabilities of key records being population uniques. Let $f_1, \dots, f_K$ denote the sample cell counts in the contingency table of key variables and $F_1, \dots, F_K$ the corresponding population cell counts and let n and N represent the sample and population totals respectively. Skinner and Elliot (2002) discuss three possible measures of risk

$$P(PU) = \frac{1}{N}\sum_{i=1}^{K} I(F_i = 1) \tag{1}$$

†*Address for correspondence:* Southampton Statistical Sciences Research Institute, School of Mathematics, University of Southampton, Highfield, Southampton, SO17 1BJ, UK.
E-mail: J.J.Forster@soton.ac.uk.

the probability of a randomly chosen record being a population unique,

$$P(PU|SU) = \frac{\sum_{i=1}^{K} I(F_i = 1, f_i = 1)}{\sum_{i=1}^{K} I(f_i = 1)} \tag{2}$$

the probability of a randomly chosen sample unique being a population unique, and

$$\theta = \frac{\sum_{i=1}^{K} I(f_i = 1)}{\sum_{i=1}^{K} F_i I(f_i = 1)} \tag{3}$$

the probability of a randomly chosen population record from a sample unique cell being the actual sampled record, where, in each case, $I(\cdot)$ is an indicator function taking the value 1 if $\cdot$ is true and 0 otherwise. Skinner and Elliot (2002) argue that θ gives the most appropriate measure of overall disclosure risk. For Bernoulli sampling, where each population record is sampled independently with common probability π, they propose the design-based estimator

$$\hat{\theta} = \frac{\pi n_1}{\pi n_1 + 2(1-\pi) n_2} \tag{4}$$

where n_1 is the number of sample uniques and n_2 is the number of sample cells with frequency 2. They show that, under a particular asymptotic framework, $\hat{\theta}$ is a consistent estimator of θ.

The above quantities all measure the overall disclosure risk of the data release. However, it is also important to be able to assess the risk associated with the release of individual records. Benedetti and Franconi (1998) noted that the probability of a randomly chosen population record from a particular cell i matching a selected sampled record is $1/F_i$, and suggested that this quantifies the disclosure risk for cell i. Alternatively, one might consider risk to be defined by uniqueness, and hence quantify risk by the binary indicator $I[F_i = 1, f_i = 1]$ identifying those sample unique cells i which are also population uniques.

Typically, the contingency tables involved are large. Skinner and Holmes (1998) present an example with of a 6-dimensional table with 83 600 cells, Polettini and Stander (2005) and Elamir and Skinner (2005) use 5-dimensional tables with 20 384 and 14 560 cells, respectively. The tables are typically also very sparse. In the example of Skinner and Holmes (1998) 74 067 cells are empty in the population and 79 603 are empty in the sample. The example we use in this paper is of a similar size and structure; see Section 5 for details. The functions of interest focus on cells with low sample and population counts. Hence, the estimation problem is different from more standard contingency table inference problems, where interest is more likely to be focussed on the cells with larger frequencies.

In this paper, we propose a model-based Bayesian approach to estimation of such measures of disclosure risk. This approach has elements in common with other approaches to the uniques problem, which we summarise here. Bethlehem et al (1990) proposed a Poisson-Gamma superpopulation model where population cell frequencies F_i are modelled as independent Poisson observations with means $N\pi_i$. The π_i parameters are given independent gamma distributions with parameters α and β, which are estimated using the sample cell frequencies f_i. It is then straightforward to estimate, for example, $P(PU|SU)$ (or equivalently under this model $I[F_i = 1, f_i = 1]$) using $[(1 + n\hat{\beta})/(1 + N\hat{\beta})]^{\hat{\alpha}+1}$; see Rinnott (2003). The Poisson-gamma model is slightly unnatural, as the F_i should be constrained to sum to N, and the π_i to sum to one. Takemura (1999) proposed an equivalent Dirichlet-multinomial model which incorporates these constraints; see also Polettini and Stander (2005).

One way of interpreting the approach described above, where the superpopulation parameters α and β are estimated and subsequent inference is based on these estimates is as parametric empirical Bayes estimation, where the Poisson-gamma (superpopulation) distribution is considered as a prior distribution, but its parameters are estimated based on the marginal likelihood of the data, rather than specified. Omori (1999) considers the Dirichlet-multinomial distribution of Takemura (1999) as a prior for the population and provides expressions for posterior probabilities of population uniqueness, given sample uniqueness. Rinott (2003) shows that, by imposing α and β with $\alpha\beta = 1/K$ and $\alpha \to 0$ in the model of Bethlehem et al (1990), we obtain the model of Benedetti and Franconi (1998), where the distribution for the $F_i|f_i$ are assumed to be independent negative binomial.

All of the models described above generally treat the cells of the underlying contingency table as exchangeable. Hence, for example, any two sample uniques will be estimated to have the same probability of being population unique under these models. Similarly, although the models allow 'borrowing strength' across cells, this is restricted to a shrinkage towards a uniform pattern of population frequencies. In particular, each other cell in the table contributes the same amount to the inference concerning any given cell. Intuitively, we might expect closer cells, in the sense of having more values of the cross-classifying variables in common, to be weighted more highly in this process. Hence, although borrowing strength is desirable, particularly in sparse tables, we would expect it to be achieved in a way which reflects the natural (multivariate) structure of the data.

One approach which does precisely this was proposed by Skinner and Holmes (1998). They replaced the Poisson-gamma superpopulation (prior) distribution with a Poisson-lognormal distribution. The population cell frequencies F_i are modelled as independent Poisson observations with means λ_i which are given independent lognormal distributions. However, the means of the lognormal distributions for the cells follow a log-linear model. This does not imply that the population frequencies reflect the given model, as the variance of the lognormal may be large. Hence, the superpopulation parameters consist of the log-linear parameters together with a lognormal variance. This model does not treat the cells of the underlying contingency table as exchangeable, and allows differential borrowing of strength from other cells. Skinner and Holmes approach to estimation is analogous to that of Bethlehem et al (1990), in that the log-linear parameters are estimated using the sample frequencies. Elamir and Skinner (2004) adapt this approach by dropping the second-stage lognormal distribution and assuming that the population follows the given log-linear model. Computation is then more straightforward.

An alternative approach, also based on a log-linear model, was proposed by Fienberg and Makov (1998). They propose assessing disclosure risk by considering a probability distribution for the $N - n$ non-released records based on the conditional distribution of a table of size n, under a particular log-linear model, given the sufficient statistics for the model parameters. Specifically, the non-released records are considered to have been generated as the sum of a set of $(N - n)/n$ tables of size n each with the same sufficient statistics as the released table. Although we do not adopt this approach in the current paper, we note that it is again based on a particular log-linear model, and Fienberg and Makov (1998) make the important point that, with such a method, the potential for model mis-specification becomes an issue. Skinner and Holmes (1998) find that their results can be sensitive to the log-linear model they choose for their λ_i parameters. Fienberg and Makov (1998) suggest that model uncertainty could be taken into account by using a Bayesian approach, and it is such an approach, but applied to the Multinomial-Dirichlet framework of Omori (1999), that we adopt in this paper.

The key feature of our approach is to allow for model uncertainty, in a fashion which is much more explicit than in the approaches described above. There, lack of model fit is accounted for, if at all, by lognormal (or gamma) variation. There are potential estimation benefits in considering a multiplicity of models, and using the sample information to determine which are supported by the data, and hence are likely to be of use for disclosure risk estimation. Furthermore, the Bayesian approach we adopt does not require a single model from a candidate set to be selected for the purposes of prediction, and allows model uncertainty to be fully incorporated into estimation. In the following sections we describe the theory behind the approach, and develop efficient methods for computing the resulting disclosure risk summaries. An additional attractive feature of the Bayesian approach adopted is that these summaries can all be easily interpreted as predictive probabilities for some disclosure event.

2. Bayesian inference

In our Bayesian approach, the population cell frequencies F_i are considered as the main parameters (unknowns). Inference about these parameters will enable inference about the various measures of disclosure risk discussed in Section 1. Therefore, we require a prior for these cell frequencies. Following Omori (1999), we assume that $\boldsymbol{F} = (F_1, \ldots, F_K)$ has a multinomial$(N, \boldsymbol{\pi})$ distribution. Now, we assume a log-linear model for $\boldsymbol{\pi}$,

$$\log \boldsymbol{\pi} = \boldsymbol{X}_m \boldsymbol{\beta}_m$$

where m indexes the particular model under consideration, and $\boldsymbol{X}_m$ and $\boldsymbol{\beta}_m$ are the corresponding model matrix and vector of model parameters, respectively. The prior may be completed by adding a further hierarchy, consisting of a prior distribution for the log-linear parameters $\boldsymbol{\beta}_m$. One possibility is a multivariate normal distribution, but this leads to an intractable posterior distribution, which generally requires Markov chain Monte Carlo computation. Nevertheless, for general log-linear models, no alternative tractable prior exists. However, if a log-linear model m is a *decomposable graphical model*, then the hyper-Dirichlet family, a class of prior distributions based on the Dirichlet distribution for the saturated model (no log-linear constraints) and developed by Dawid and Lauritzen (1994), provides an attractive alternative, for which posterior computation is straightforward.

For details of graphical models, see Cowell *et al* (1999). Briefly, an undirected graphical model can be represented by a graph where the vertices correspond to variables, and the absence of an edge between two vertices represents conditional independence of the corresponding variables, given the values of the other variables. A decomposable undirected graph is one with no chordless cycles of four or more vertices (in other words, every circuit of four or more vertices, has an edge 'short-cutting' the circuit). A decomposable graph admits a *perfect numbering*, that is a numbering $1, \ldots, v$ of the vertices, such that for each vertex number $i = 2, \ldots, v$, the set of those vertices numbered in $\{1, \ldots, i-1\}$ which share an edge with i in the graph is complete (there are no missing edges in the graph formed by this subset). An important consequence for Bayesian analysis is that, for a decomposable graphical model we can exploit the alternative parameterisation

$$\pi_i(\boldsymbol{\beta}_m) = \prod_{j=1}^{v} \beta_j(i_j | \text{pa}(j) = i_{\text{pa}(j)}) \tag{5}$$

where β_j is the conditional probability for variable j and the ordering $1,\ldots,v$ of the variables corresponds to a perfect numbering. The index i_C represents the cell of the marginal table for variable set C corresponding to cell i for the full table and, for any variable j, $\text{pa}(j)$ represents the subset of $\{1,\ldots,j-1\}$ which share an edge with j in the graph of m. The probability vectors $\boldsymbol{\beta}_j(i_{\text{pa}(j)}) = \{\beta_j(i_j|\text{pa}(j) = i_{\text{pa}(j)}), i_j \in \text{levels of } j\}$ are then unconstrained by the model, and hence parameterise the model, so we write $\boldsymbol{\beta}_m = (\{\boldsymbol{\beta}_j(i_{\text{pa}(j)})\}, j = 1,\ldots,v)$ where $\{\boldsymbol{\beta}_j(i_{\text{pa}(j)})\}$ is the collection of conditional probability vectors for variable j over all combinations of levels of $\text{pa}(j)$.

A convenient prior distribution for the unconstrained probability vectors $\boldsymbol{\beta}_m$ can be constructed by putting independent Dirichlet$[\boldsymbol{\mu}_j(i_{\text{pa}(j)})]$ prior distributions on each $\boldsymbol{\beta}_j(i_{\text{pa}(j)})$. Then the posterior distributions are independent Dirichlet$[\boldsymbol{\mu}^*_j(i_{\text{pa}(j)})]$ where

$$\boldsymbol{\mu}^*_j(i_{\text{pa}(j)}) = \boldsymbol{\mu}_j(i_{\text{pa}(j)}) + \boldsymbol{f}_j(i_{\text{pa}(j)}) \tag{6}$$

and $\boldsymbol{f}_j(i_{\text{pa}(j)})$ is the vector of marginal cell counts for variable j derived from the conditional subtable restricted to $\text{pa}(j) = i_{\text{pa}(j)}$. The hyper-Dirichlet distribution can be thought of as a special case of this independent Dirichlet prior, where the $\boldsymbol{\mu}_j(i_{\text{pa}(j)})$ are chosen in a way such that the resulting prior distribution on the cell probabilities for the clique margins (marginal tables which are unrestricted by the model) are Dirichlet. This is easy to achieve if, for example, the $\boldsymbol{\mu}_j(i_{\text{pa}(j)})$ are derived as marginal distributions from a common Dirichlet distribution over $\boldsymbol{\pi}$; see Dawid and Lauritzen (1994) for further details.

Where the number of cross-classifying variables is small (up to 3 or 4, say) the restriction to decomposable graphical models is unlikely to be significant, as most hierarchical log-linear models are also decomposable graphical models. For example, for three variables, we only lose the *no three-way interaction* model. In higher dimensions, the model class is considerably reduced, but still provides a rich class of structural models with which we might expect to considerably improve upon any disclosure assessment approach which assumes an exchangeable model.

Having specified a prior distribution, and observed sample cell frequencies $\boldsymbol{f} = (f_1,\ldots,f_K)$, inference concerning disclosure risk is obtained from the posterior distribution $P(\boldsymbol{F}|\boldsymbol{f})$ or equivalently from $P(\boldsymbol{F}-\boldsymbol{f}|\boldsymbol{f})$, as having observed the sampled records it is only the unsampled records about which uncertainty remains. Assuming a sampling scheme under which records are exchangeable, for example Bernoulli sampling or simple random sampling without replacement, then we can assume that $\boldsymbol{F}-\boldsymbol{f}$ and $\boldsymbol{f}$ are conditionally independent given $\boldsymbol{\beta}_m$ and hence, following Ericson (1969), we can write

$$P(\boldsymbol{F}-\boldsymbol{f}|\boldsymbol{f}) = \int P(\boldsymbol{F}-\boldsymbol{f}|\boldsymbol{\beta}_m)P(\boldsymbol{\beta}_m|\boldsymbol{f})\mathrm{d}\boldsymbol{\beta}_m. \tag{7}$$

The first term in the integrand, $P(\boldsymbol{F}-\boldsymbol{f}|\boldsymbol{\beta}_m)$, is just a multinomial$(N-n,\boldsymbol{\pi})$ probability function, with probabilities $\boldsymbol{\pi}$ determined from $\boldsymbol{\beta}_m$ using (5). For marginal inference about a single cell, $P(\boldsymbol{F}-\boldsymbol{f}|\boldsymbol{\beta}_m)$ is replaced by the binomial probability function $P(F_i-f_i|\boldsymbol{\beta}_m)$. Using Bayes theorem, the second term of the integrand is

$$P(\boldsymbol{\beta}_m|\boldsymbol{f}) \propto P(\boldsymbol{f}|\boldsymbol{\beta}_m)P(\boldsymbol{\beta}_m),$$

the product of a multinomial$(n,\boldsymbol{\pi})$ probability function for $\boldsymbol{f}$ and the prior density for $\boldsymbol{\beta}_m$. The sample records may have been selected using a more complex sampling scheme. However, a potential intruder's knowledge of the sampling framework, and its relationship to the sampled and unsampled units is still likely to be sufficiently weak that a prior distribution which is exchangeable, with respect to individual records, is appropriate.

To this point, we have only described inference under a single log-linear model. In practice, it is unlikely that we will be certain about which model is the most appropriate for building the prior distribution for $\boldsymbol{F}$. A Bayesian approach allows this uncertainty to be coherently incorporated into the prior distribution. Let M denote the set of possible models, and suppose that prior uncertainty about m is encapsulated by a prior distribution over M, involving a set of prior model probabilities $P(m)$. In practice, a discrete uniform distribution over M is commonly used, to represent prior ignorance. The prior distribution over $\boldsymbol{F}, m$ and $\{\boldsymbol{\beta}_m, m \in M\}$ now consists of three components, the multinomial $P(\boldsymbol{F}|\boldsymbol{\beta}_m, m)$, the prior for the parameters of each possible log-linear model $P(\boldsymbol{\beta}_m|m)$ and the prior model probabilities $P(m)$. Note that the first two distributions are now explicitly conditional on m, as both the form of the log-linear model, and the prior distribution for its parameters, will depend on which model is under consideration.

Under model uncertainty, the posterior distribution for the unobserved cell counts $\boldsymbol{F}-\boldsymbol{f}$ in (7) becomes

$$P(\boldsymbol{F}-\boldsymbol{f}|\boldsymbol{f}) = \sum_{m\in M} P(m|\boldsymbol{f}) \int P(\boldsymbol{F}-\boldsymbol{f}|N-n, \boldsymbol{\beta}_m, m) P(\boldsymbol{\beta}_m|\boldsymbol{f}, m) \mathrm{d}\boldsymbol{\beta}_m. \tag{8}$$

The posterior model probabilities, which appear in (8) but not (7) are obtained, using Bayes theorem as

$$P(m|\boldsymbol{f}) = \frac{P(m)P(\boldsymbol{f}|m)}{\sum_{m\in M} P(m)P(\boldsymbol{f}|m)} \tag{9}$$

where $P(\boldsymbol{f}|m)$ is the marginal likelihood for the sampled cell counts, obtained as

$$P(\boldsymbol{f}|m) = \int P(\boldsymbol{f}|m, \boldsymbol{\beta}_m) P(\boldsymbol{\beta}_m|m) \mathrm{d}\boldsymbol{\beta}_m. \tag{10}$$

The posterior distribution (8) under model uncertainty is obtained as a weighted average of the posterior distributions (7) under the various models. This is sometimes referred to as *model-averaging*; see Hoeting *et al* (1999) for further details. Where the quantity of interest is a posterior expectation for some function $\theta(\boldsymbol{F})$, as are all the measures of disclosure risk considered in this paper, then we obtain

$$E[\theta(\boldsymbol{F})|\boldsymbol{f}] = \sum_{m\in M} P(m|\boldsymbol{f}) E[\theta(\boldsymbol{F})|m, \boldsymbol{f}]. \tag{11}$$

and hence the posterior mean taking into account model uncertainty is simply the model-average of the posterior means under each of the models. Care is required when performing model-averaging, that the quantity which is being averaged is one which shares a common interpretation across the component models. That is clearly the case here, where we are averaging functions of cell frequencies. The posterior model probabilities are not of particular interest in themselves here, as we do not necessarily believe that the population was exactly generated by a particular multinomial log-linear model. Their function is to indicate the appropriate weight, based on the sample data, to be applied to the various models in any inference required. Consequently, they determine the differential impact of other cells, when making inference about a particular population cell frequency.

The overall measures of disclosure risk in (1)-(3) and $I[F_i = 1, f_i = 1]$ and $1/F_i$ as measures of record-level risk, each depend on the population frequencies $\boldsymbol{F}$. Indeed they can each be thought of as $P(E|\boldsymbol{F})$ where E represents an event indicative of disclosure risk. If the population frequencies $\boldsymbol{F}$ are known, then $P(E|\boldsymbol{F})$ is a single known probability, but given only the sample data $\boldsymbol{f}$, uncertainty exists about the true value of any $P(E|\boldsymbol{F})$. This uncertainty can be expressed through the posterior probability distribution. Posterior inference is straightforward in principle. The marginal posterior distribution of $P(PU)$, $P(PU|SU)$ or θ can be obtained directly from the posterior distribution $P(\boldsymbol{F}-\boldsymbol{f}|\boldsymbol{f})$ in (7) or (8). For example,

$$P(\theta = \phi|\boldsymbol{f}) = \sum_{\boldsymbol{F}\in\mathcal{F}} P(\boldsymbol{F}|\boldsymbol{f}) \tag{12}$$

where $\mathcal{F}$ is the set of those $\boldsymbol{F}$ such that the probability of interest, θ, takes the value ϕ.

We note that $P(E|\boldsymbol{f}) = E[P(E|\boldsymbol{F})|\boldsymbol{f}]$ where the expectation is with respect to the posterior distribution $P(\boldsymbol{F}-\boldsymbol{f}|\boldsymbol{f})$. Hence, the posterior mean of any $P(E|\boldsymbol{F})$ is the predictive probability of event E, given the observed sample data alone representing, from a Bayesian perspective, the probability of E given the available information. For example, θ is the probability of matching a uniform randomly selected population record in a sample unique cell to a sample unique over potential multiple attempts, sampling with replacement. The posterior distribution for θ reflects our uncertainty about that (frequentist) probability. However, the posterior mean of θ is a particularly important measure of disclosure risk, as it represents the predictive probability, given the available data, of matching a randomly selected population record to a sample unique in a single attempt.

Hence, we focus on posterior means of $P(PU)$, $P(PU|SU)$ or θ to summarise overall disclosure risk. In the case of the first two quantities, these posterior means can be simply expressed as

$$\begin{aligned} E[P(PU)|\boldsymbol{f}] &= \frac{1}{N}\sum_{i=1}^{K} P(F_i = 1|\boldsymbol{f}) \\ &= \frac{1}{N}\sum_{i=1}^{K} P(F_i - f_i = 1|\boldsymbol{f})I(f_i = 0) \\ &\qquad +\frac{1}{N}\sum_{i=1}^{K} P(F_i - f_i = 0|\boldsymbol{f})I(f_i = 1) \end{aligned} \tag{13}$$

and

$$E[P(PU|SU)|\boldsymbol{f}] = \frac{\sum_{i=1}^{K} P(F_i - f_i = 0|\boldsymbol{f})I(f_i = 1)}{\sum_{i=1}^{K} I(f_i = 1)} \tag{14}$$

respectively. Similarly, at the record level

$$E[1/F_i|\boldsymbol{f}] = \sum_{j=0}^{N-n} \frac{1}{f_i + j} P(F_i - f_i = j|\boldsymbol{f}) \tag{15}$$

and

$$E[I[F_i = 1, f_i = 1]|\boldsymbol{f}] = P(F_i - f_i = 0|\boldsymbol{f})I[f_i = 1] \tag{16}$$

Hence, each of these disclosure risk measures may be represented as explicit functions of predictive probabilities $P(\boldsymbol{F}-\boldsymbol{f}|\boldsymbol{f})$. This is not possible for $E[\theta|\boldsymbol{f}]$ where θ is defined by (3).

To compute this disclosure risk measure it is necessary first to enumerate, or approximate the posterior distribution $P(\theta|\boldsymbol{f})$ as in (12).

3. Monte Carlo Computation

There are three computational difficulties associated with calculating the predictive probabilities which are proposed as disclosure risk measures. The first is the evaluation of the integrals in (7) and (8). These integrals are analytically intractable for general log-linear models, where the prior on the log-linear parameters is multivariate normal, as suggested in Section 2. However, they can be straightforwardly evaluated when a hyper-Dirichlet prior distribution is used for a decomposable log-linear model. The second problem is evaluation of the sum in (8), in cases where the number of models is so large that evaluation of the summand for every model is infeasible. Finally, evaluation of the sum in (12) can also be prohibitively time consuming, as it is over a potentially large space of possible values of (multivariate) $\boldsymbol{F}$. Exactly the same consideration applies when computing the marginal posterior distributions of individual population cell frequencies $P(F_i|\boldsymbol{f})$ required in (13)-(16), as the marginal predictive probabilities $P(F_i - f_i|\boldsymbol{f})$ are not directly available for a hyper-Dirichlet posterior and hence must be enumerated using $P(\boldsymbol{F}|\boldsymbol{f})$.

Monte Carlo methods of computation are particularly attractive as, given a sample from the posterior distribution of $\boldsymbol{F}$, or equivalently $\boldsymbol{F} - \boldsymbol{f}$, the probabilities $P(F_i|\boldsymbol{f})$ are simply estimated by sample proportions, which can then be plugged into (13)-(16). Similarly $E[\theta|\boldsymbol{f}]$ can be approximated by a Monte Carlo sample average. To sample from the posterior distribution of $\boldsymbol{F}$, we sample from the joint posterior distribution $P(m, \boldsymbol{\beta}_m|\boldsymbol{f})$ and then the multinomial $P(\boldsymbol{F}|m, \boldsymbol{\beta}_m)$. As we are assuming that $\boldsymbol{F} - \boldsymbol{f}$ and $\boldsymbol{f}$ are conditionally independent given m and $\boldsymbol{\beta}_m$, then

$$P(F_i - f_i|\boldsymbol{f}) = E[P(F_i - f_i|m, \boldsymbol{\beta}_m)|\boldsymbol{f}], \qquad i = 1, \ldots, k,$$

where the expectation is over the posterior distribution $P(m, \boldsymbol{\beta}_m|\boldsymbol{f})$. Hence, more accurate, 'Rao-Blackwellised' estimates of the probabilities $P(F_i - f_i|\boldsymbol{f})$ are obtained by averaging the binomial probabilities

$$P(F_i - f_i|m, \boldsymbol{\beta}_m) = \binom{N-n}{F_i - f_i} \pi_i^{F_i - f_i} (1 - \pi_i)^{N-n-F_i+f_i} \tag{17}$$

where the cell probabilities $\boldsymbol{\pi}$ are determined from m and $\boldsymbol{\beta}_m$ using (5), over a sample generated from the joint posterior distribution of m and $\boldsymbol{\beta}_m$. Hence, more accurate Monte Carlo estimates of quantities such as (13)-(16) are available, without ever sampling $\boldsymbol{F}$, by replacing $P(F_i - f_i|\boldsymbol{f})$ by $E[P(F_i - f_i|m, \boldsymbol{\beta}_m)|\boldsymbol{f}]$, estimated using Monte Carlo sample averages of (17).

Computing the posterior mean of θ, using the marginal posterior distribution (12), requires a summation over all possible sets of population cell frequencies $\boldsymbol{F}$, and hence Monte Carlo estimation of θ is only feasible by generating a sample of $\boldsymbol{F}$ from its posterior distribution. Similar considerations may apply to (15) where the number of terms in the summand $(N - n + 1)$ is typically large, leading Monte Carlo generation of $\boldsymbol{F}$ to be the preferred method of calculation.

For decomposable models and hyper-Dirichlet prior distributions, some of the calculations can be performed exactly. For example, marginal likelihoods (10) may be computed explicitly, as

$$P(\boldsymbol{f}|m)=\prod_{j=1}^{v}\prod_{i_{\mathrm{pa}(j)}}\frac{\Gamma\left[\sum_{i_j}\mu_j(i_j|i_{\mathrm{pa}(j)})\right]\prod_{i_j}\Gamma[\mu_j^*(i_j|i_{\mathrm{pa}(j)})]}{\Gamma\left[\sum_{i_j}\mu_j^*(i_j|i_{\mathrm{pa}(j)})\right]\prod_{i_j}\Gamma[\mu_j(i_j|i_{\mathrm{pa}(j)})]} \tag{18}$$

where $\mu_j(i_j|i_{\mathrm{pa}(j)})$ represents the i_jth component of $\boldsymbol{\mu}_j(i_{\mathrm{pa}(j)})$, and $\boldsymbol{\mu}_j^*(i_{\mathrm{pa}(j)})$ is given by (6). Hence, provided that the number of models under consideration is not too great, posterior model probabilities (9) are available. For more than a few (3 or 4) cross-classifying variables, it is unlikely to be feasible to calculate posterior probabilities for all models, and some kind of Monte Carlo sampling may be required. Madigan and York (1996) describe an efficient Markov chain Monte Carlo (MCMC) approach to sampling decomposable graphical models according to their posterior model probabilities.

Having obtained model probabilities, the posterior distributions for the model parameters $\boldsymbol{\beta}_j(i_{\mathrm{pa}(j)})$ for any model m are independent Dirichlet$[\boldsymbol{\mu}_j^*(i_{\mathrm{pa}(j)})]$. Now, $P(F_i-f_i|m,\boldsymbol{\beta}_m)$ in (17) involves a polynomial of degree $N-n$ in the cell probabilities. For the product-Dirichlet distribution, this necessitates expanding the polynomial and evaluating the expectation of each of up to $N-n+1$ monomial terms, for which exact product-Dirichlet expectations are available, as

$$\begin{aligned} E[\pi_i^t] &= \prod_{j=1}^{v}E[\beta_j(i_j|i_{\mathrm{pa}(j)})^t] \\ &= \prod_{j=1}^{v}\frac{\Gamma\left[\sum_{k_j}\mu_j^*(k_j|i_{\mathrm{pa}(j)})\right]\Gamma[\mu_j^*(i_j|i_{\mathrm{pa}(j)})+t]}{\Gamma\left[\sum_{k_j}\mu_j^*(k_j|i_{\mathrm{pa}(j)})+t\right]\Gamma[\mu_j^*(i_j|i_{\mathrm{pa}(j)})]}. \end{aligned} \tag{19}$$

Where $N-n$ is large, and interest is focussed on small values of F_i-f_i, as is typically the case, it is more efficient to estimate the posterior mean of (17) by Monte Carlo. This involves generating from the independent Dirichlet distributions for $\boldsymbol{\beta}_j(i_{\mathrm{pa}(j)})$ and substituting into (5) to obtain a Monte Carlo sample from $P(\boldsymbol{\pi}|\boldsymbol{f})$. Similarly, computation of $E(\theta|\boldsymbol{f})$ using (12) can be carried out by generating $\boldsymbol{F}-\boldsymbol{f}$ from a multinomial$(N-n,\boldsymbol{\pi})$ for each $\boldsymbol{\pi}$ generated from $P(\boldsymbol{\pi}|\boldsymbol{f})$.

4. Fast approximations

Estimating (13)-(16) by Monte Carlo methods can be time consuming, particularly for tables with large numbers of cells, as is typical in disclosure applications. In such examples, it is desirable to try and identify approximations which avert the requirement for generating samples from posterior distributions.

The Monte Carlo calculations in Section 3 are primarily used to estimate $E[P(F_i-f_i|m,\boldsymbol{\beta}_m)|\boldsymbol{f}]$ using Monte Carlo averages of (17) over samples generated from $P(m,\boldsymbol{\beta}_m|\boldsymbol{f})$. In the examples we are interested in, $N-n$, which we now denote by u, is large so we replace (17) with the Poisson approximation

$$P(F_i-f_i|m,\boldsymbol{\beta}_m)\approx\frac{(u\pi_i)^{F_i-f_i}\exp(-u\pi_i)}{(F_i-f_i)!} \tag{20}$$

Under the hyper-Dirichlet distribution for $\boldsymbol{\beta}_m$ associated with each model m, the distribution for cell probabilities π_i can be expressed as a product of v independent beta variates.

Hence, it is possible to evaluate $E(\pi_i|\boldsymbol{f})$ and $Var(\pi_i|\boldsymbol{f})$ directly, using (19) with $t = 1$ and $t = 2$. We propose approximating the distribution of $\pi_i|\boldsymbol{f}$ by a gamma distribution with mean $E(\pi_i|\boldsymbol{f})$ and variance $Var(\pi_i|\boldsymbol{f})$. This approximation is certainly plausible as in the special case of the saturated model (no conditional independence structure), the marginal distribution of $\pi_i|\boldsymbol{f}$ is beta with parameters whose sum is the sample size plus the sum over all cells of the prior parameters. The beta(α_1, α_2) distribution has a gamma limit, both as $\alpha_2 \to \infty$ for fixed α_1, and as $\alpha_1, \alpha_2 \to \infty$ at the same rate. Hence, with a sufficiently large sample, a gamma approximation is valid for $\pi_i|\boldsymbol{f}$, for cells i with both small and moderate posterior means. Fan (1991) provides numerical results which suggest that the product of independent betas may be approximated by a beta distribution with the same moments. Our approach then replaces this approximating beta(α_1, α_2) distribution with a gamma distribution, which is justified above, provided that α_2 is large.

A gamma(α, λ) approximation for the posterior distribution of $\pi_i|\boldsymbol{f}$, together with the Poisson approximation (20) leads to the negative binomial approximation

$$P(F_i - f_i|\boldsymbol{f}) \approx \left(\frac{\lambda}{\lambda+u}\right)^{\alpha} \left(\frac{u}{\lambda+u}\right)^{F_i-f_i} \frac{\Gamma(\alpha+F_i-f_i)}{\Gamma(\alpha)(F_i-f_i)!} \tag{21}$$

For the *saturated* model (no log-linear simplification) this is equivalent to the negative binomial distribution used by Bethlehem et al (1990), as presented by Rinnott (2003). As Rinnott (2003) shows, this is also a special case of the distribution proposed by Benedetti and Franconi (1998). For non-saturated models, the approximations for α and λ lead to negative binomial distributions whose parameters differ from previous approaches.

The expression in (21) simplifies in (13), (14) and (16) to

$$P(F_i - f_i = 0|\boldsymbol{f}) \approx \left(\frac{\lambda}{\lambda+u}\right)^{\alpha} \tag{22}$$

and

$$P(F_i - f_i = 1|\boldsymbol{f}) \approx \left(\frac{\lambda}{\lambda+u}\right)^{\alpha} \frac{\alpha u}{\lambda+u} \tag{23}$$

Furthermore, (15) can now be written as

$$\begin{aligned} E[1/F_i|\boldsymbol{f}] &\approx \sum_{j=0}^{\infty} \frac{1}{f_i+j} \left(\frac{\lambda}{\lambda+u}\right)^{\alpha} \left(\frac{u}{\lambda+u}\right)^{F_i-f_i} \frac{\Gamma(\alpha+F_i-f_i)}{\Gamma(\alpha)(F_i-f_i)!} \\ &= \left(\frac{\lambda}{\lambda+u}\right)^{\alpha} \frac{1}{f_i}\ {}_2F_1(\alpha, f_i, f_i+1, u/(\lambda+u)) \end{aligned} \tag{24}$$

where ${}_2F_1$ is the hypergeometric function (see, for example, Abramowitz and Stegun, 1965), which can be computed in R, as part of the Davies package (Hankin, 2005). Polettini (2003) notes that a similar expression arises in the approach of Benedetti and Franconi (1998), where they are required to calculate $E(1/F_i|f_i)$, under the assumption of a negative binomial model for $F_i|f_i$.

Finally, an approximation to $E[\theta|\boldsymbol{f}]$ where θ is defined by (3) can be obtained by approximating the distribution of $\sum_{i=1}^{K}(F_i - f_i)I(f_i = 1)|m, \boldsymbol{\beta}_m$ as a Poisson distribution with mean $u\sum_{i=1}^{K} \pi_i I(f_i = 1)$ and the posterior distribution of $\sum_{i=1}^{K} \pi_i I(f_i = 1)$ by a gamma(α, λ) distribution where α and λ are derived from the means and variances of the individual π_i, as if they were independent (which they are not). Then,

$$E[\theta|\boldsymbol{f}] \approx \left(\frac{\lambda}{\lambda+u}\right)^{\alpha} \frac{1}{f}\ {}_2F_1(\alpha, f, f+1, u/(\lambda+u)) \tag{25}$$

where $f = \sum f_i I(f_i = 1)$. If $f\lambda/u \gg 1$, then this approximation may be simplified to $E[\theta|\boldsymbol{f}] \approx f/(f + u\alpha/\lambda)$. This approximation is somewhat more speculative than those in (22)-(24). The performance of all these approximations will be investigated in our example in Section 5

The approximations discussed here are all based on a single decomposable graphical model. In the presence of model uncertainty, the computations are model-averaged, as in (8) and (11), and we desire a method for evaluating or approximating the posterior model probabilities $P(m|\boldsymbol{f})$ which avoids Monte Carlo sampling. Our approach is based on the 'Occam's window' method of Madigan and Raftery (1994). Starting from the null (complete independence) model, we identify a set of 'most plausible' models, as follows. Our current set of models is augmented to include all models with exactly one extra edge. The maximum probability model, m_0, in this set is identified. Then, all models, m_i, with $P(m_i|\boldsymbol{f})/P(m_0|\boldsymbol{f})$ less than some prespecified (small) threshold are deleted. The procedure is iterated, starting from the null (complete independence) model until the model set remains unchanged. As we are using hyper-Dirichlet priors, relative posterior model probabilities are very conveniently calculated (see Madigan and Raftery, 1994). The final model probabilities are calculated under the assumption that any model outside our final set has zero posterior probability, and it is here that the calculation is only approximate. The final computed disclosure risk estimated is obtained by computing the relevant quantity for each model in our final set, and model-averaging with respect to the approximate posterior model probabilities.

Our entire approach to computing disclosure risk estimates can now be summarised as follows. Using the released sample data, we apply the search strategy described above to determine a set S of plausible decomposable graphical models, with associated model probabilities calculated using (9) and (18). For each model in S, all of which can be decomposed as described in (5), we then calculate the mean and variance of each π_i using (19) and hence obtain the parameters α and λ of a gamma approximation to the posterior distribution of π_i. These parameters can then be substituted into (22)-(24) to obtain approximate model-based probabilities of the various disclosure events. The predictive probability $E[\theta|\boldsymbol{f}]$ is computed using (25), following a further gamma approximation to the posterior distribution of the sum of π_i over the sample unique cells. Overall disclosure risk estimates are then computed by model-averaging the individual model-based estimates, using (11).

5. Example

To test the methodology on an example of realistic size, we extracted a six-way table of potential key variables from the 3% Individual Sample of Anonymized Records (SAR) for the 2001 UK Census (Office for National Statistics and University of Manchester, 2005). The table extracted consisted of 154295 individuals living in South West England, cross-classified by sex (2 categories) age (coded into 11 categories), accomodation type (8 categories), number of cars owned or available for use (5 categories), occupation type (11 categories) and family type (10 categories). The full table has 96800 cells of which 3796 are uniques. For the purposes of this exercise, this is considered to be the population. To mimic the selection

Table 1. Summary of the joint distribution of sample and population cell frequencies across the 96800 cells of the sampled table.

		Population								
		0	1	2	3	4	5-9	10-19	20+	Total
Sample	0	84867	3682	1694	967	631	1482	757	390	94470
	1	—	114	110	118	104	313	322	462	1543
	2	—	—	0	2	5	28	67	266	368
	3	—	—	—	0	0	1	15	140	156
	4	—	—	—	—	0	0	0	76	76
	5-9	—	—	—	—	—	0	0	125	125
	10-19	—	—	—	—	—	—	0	48	48
	20+	—	—	—	—	—	—	—	14	14
	Total	84867	3796	1804	1087	740	1824	1161	1521	96800

into the SAR, where the sampling fraction is 3%, we took a 3% subsample, containing 4761 individuals.

The joint distribution of sample and population cell frequencies for the 96800 cells is summarised in Table 1. Of particular note is the fact that, in the sample data, only 2330 of the 96800 cells are non-empty, and of these 1543 are uniques. Hence, 32% of records, and 66% of cells correspond to sample uniques. Of these cells, only 114 (7%) are population uniques, and the average population total in a sample unique cell is 17, so not all such cells represent disclosure risk. In any example, the number of sample unique cells will vary, depending on the total number of cells in the table of interest, and on the size of the sample, but it is not unusual for the number of sample unique cells to be large as a proportion of the number of non-empty cells. For comparison, in the actual SAR sample, 32% of the non-empty sample cells correspond to uniques; in a similar exercise, Skinner and Holmes (1998) analysed a sample with 40% unique cells.

In disclosure risk estimation, we are attempting to replicate the analysis of a potential intruder. Hence, our prior distribution should reflect the prior information available to an intruder. In the absence of any knowledge of external information available to an intruder, we propose a neutral, weakly informative, prior. Hence, we used a prior distribution where all decomposable graphical models on the 6 variables were considered *a priori* equally probable. For each model, the prior distribution on the model parameters was hyper-Dirichlet, derived from a symmetric Dirichlet distribution over all 96800 cells with parameters all equal to $p/96800$. This prior can be interpreted as having the same weight as p prior observations, and is centred on all cell probabilities being equal. Our default value for p is 1, corresponding to a single prior observation, but we also investigated sensitivity to the choice of p, by considering alternative values of $p = 50$ (equivalent to a prior sample of approximately 1% of the size of the actual observed data sample in this example) and $p = 1/50$, a highly diffuse prior. For all three prior distributions, the posterior distribution was heavily concentrated on a single model, with five edges. For that model we evaluated the accuracy of our gamma approximation to the marginal posterior distribution of the cell probabilities, in estimating our quantities of interest. For the overall measures of disclosure risk, the estimates of the posterior means of (1)-(3) obtained using both the Monte Carlo methods and approximations are displayed in Table 2 for all three prior distributions. Immediately, we notice that the approximations are very accurate, even for θ where we had least confidence that it would be.

Table 2. Estimates of the overall disclosure measures in (1)-(3) and the corresponding population value, based on $\boldsymbol{F}$.

	Estimates						Population
Measure	Monte Carlo			Gamma approximation			value
	$p = 1/50$	$p = 1$	$p = 50$	$p = 1/50$	$p = 1$	$p = 50$	
$P(PU)$	0.0256	0.0258	0.0321	0.0256	0.0257	0.0321	0.0246
$P(PU\|SU)$	0.1028	0.1027	0.0989	0.1030	0.1029	0.0991	0.0739
θ	0.0568	0.0568	0.0576	0.0568	0.0568	0.0576	0.0582

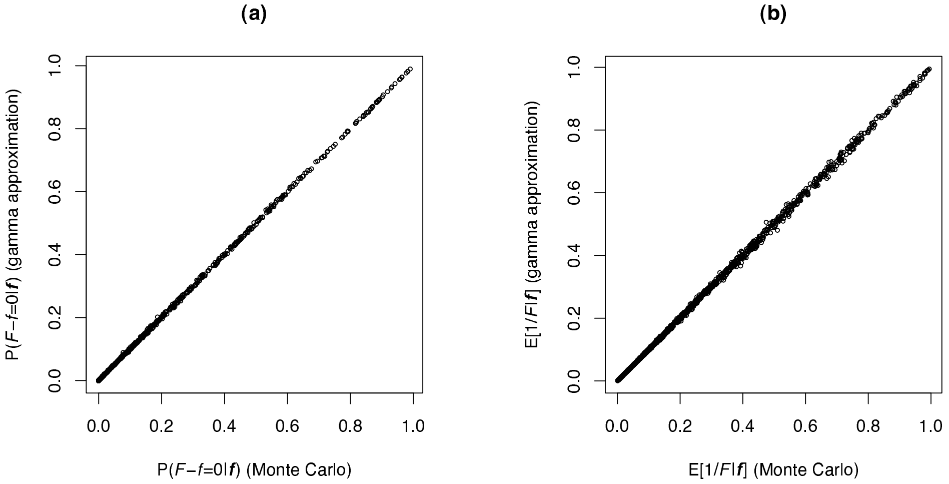

Fig. 1. Plots of the approximation based on a gamma distribution for $P(\pi_i|\boldsymbol{f})$ (y-axis) against the corresponding Monte Carlo estimate based on the true posterior distribution (x-axis) for (a) $P[F_i - f_i = 0|\boldsymbol{f}]$ (sample unique cells) and (b) $E[1/F_i|\boldsymbol{f}]$ (all cells with positive sample cell counts).

For the record-level measures of disclosure risk, obtained as the posterior means (15) and (16), we plot, for $p = 1$, the approximation against the Monte Carlo estimate in Figure 1. Again, both approximations can be seen to be accurate, with any error of approximation negligible compared with the variability of the quantities of interest across cells.

Having established that our approximations are sufficiently accurate in this example, we now address the question of how well our Bayesian methods estimate the quantities of interest. In the current example, we have the actual population available to us for comparison. For the overall measures of disclosure risk, the population values of (1)-(3) are presented in Table 2, together with the estimates. It can be seen that estimation of θ is very accurate, but $P(PU|SU)$ is overestimated. Estimation of $P(PU)$ is accurate for $p = 1/50$ and $p = 1$, but this quantity is significantly overestimated for $p = 50$. This is due to the considerably higher level of smoothing associated with this prior, resulting in the significantly larger predicted probabilities of population uniques in the sample zero cells.

As discussed in Section 2, the estimates in table 2 are the Bayesian predictive probabilities of different disclosure events, given only the sample table $\boldsymbol{f}$. Hence, they completely specify uncertainty about the corresponding disclosure events. However, if required, we could also use our approach to calculate a posterior distribution (conditional on $\boldsymbol{f}$) for the population values (the predictive probabilities we would calculate if we had the complete

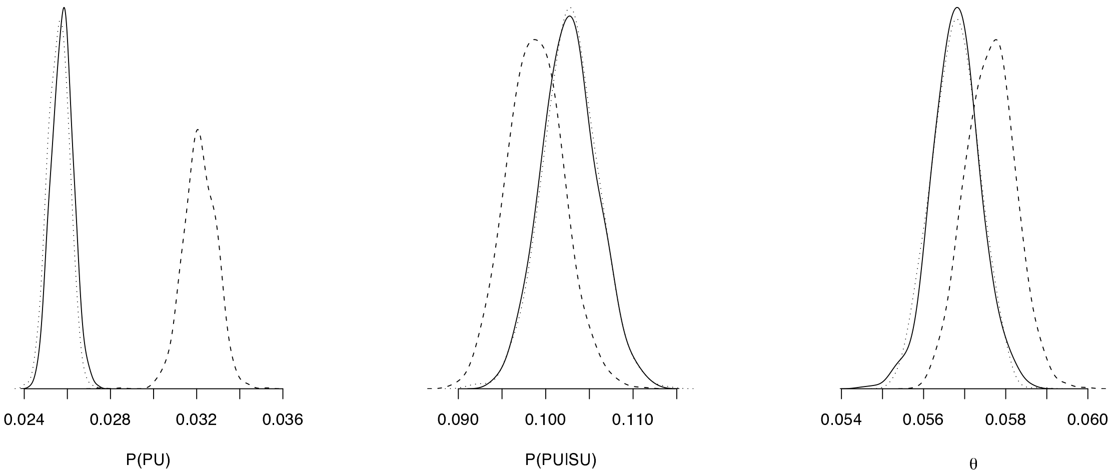

Fig. 2. The posterior densities for $P(PU)$, $P(PU|SU)$ and θ, as given in (1)-(3). Three different priors are used, $p = 1$ (solid line), $p = 50$ (dashed line) and $p = 1/50$ (dotted line, almost coincident with solid line).

population table $\boldsymbol{F}$). These posterior densities are presented, for all three priors, in Figure 2. Note that the full Monte Carlo sampling approach was required to produce these. We see that there is negligible difference between the posterior distributions for $p = 1/50$ and $p = 1$, but some sensitivity to increasing p to a value as high as 50. Again, this is most noticeable for $P(PU)$, for the reasons described above.

However, it is arguably the record-level measures of disclosure risk which are of greatest interest. Can our approach distinguish between cells with similarly low cell counts in the sample, but which pose significantly different disclosure risks? For each of the 2330 non-empty cells j, we calculated the predictive disclosure probability $E[1/F_i|\boldsymbol{f}]$. We are also able to calculate $1/F_i$, the probability of a disclosure event when full population knowledge is available. We compare these quantities, and hence assess the performance of our disclosure risk assessment procedure by plotting $\log_{10}(1/F_i)$ against the estimated $\log_{10}(E[1/F_i|\boldsymbol{f}])$ for the 2330 non-empty sample cells, in Figure 3.

Given that low frequency sample cells correspond to such a wide range of population frequencies, accurate estimation of $1/F_i$, using sample data alone is a difficult task, and without some kind of modelling would be hopeless. Indeed, using any approach which treats the cells as exchangeable, would lead, in the absence of extra external information, to all 1543 sample uniques having the same estimated risk (0.11 in this example). It is immediately clear that our approach is providing a more accurate measure of risk for the cells with low population counts (genuinely risky records). For the 114 genuine population uniques, we computed an average risk of 0.65. This risk estimate is necessarily downwardly biased, as we are considering the population units with $1/F_i$ at the boundary value of 1, but it is noticeable that the estimate does exceed the next possible population value of 0.5. For the 111 sample unique cells with population totals greater than 50, the average risk was estimated as only 0.04. Hence, the method is successfully distinguishing risky and non-risky cells with the same cell counts. In this context, our approach seems to be performing quite well, with perhaps a slight tendency to overestimate risk in this example. This slight

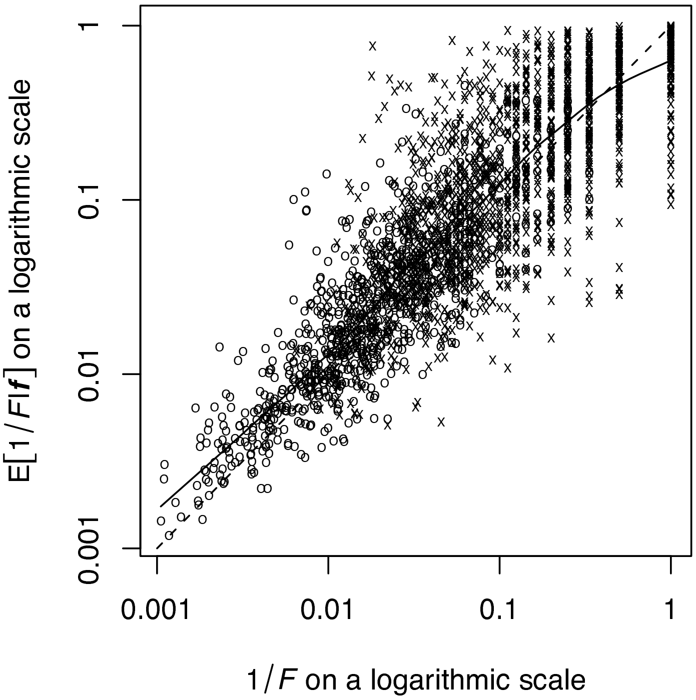

Fig. 3. The estimated record-level measure of disclosure risk $E[1/F_i|\boldsymbol{f}]$ plotted against $1/F_i$ for the 2330 non-empty sample cells. The sample unique cells are plotted as x, the non-unique cells as o. The dashed line represents equality (no error). The solid line is a loess smooth through the plotted points.

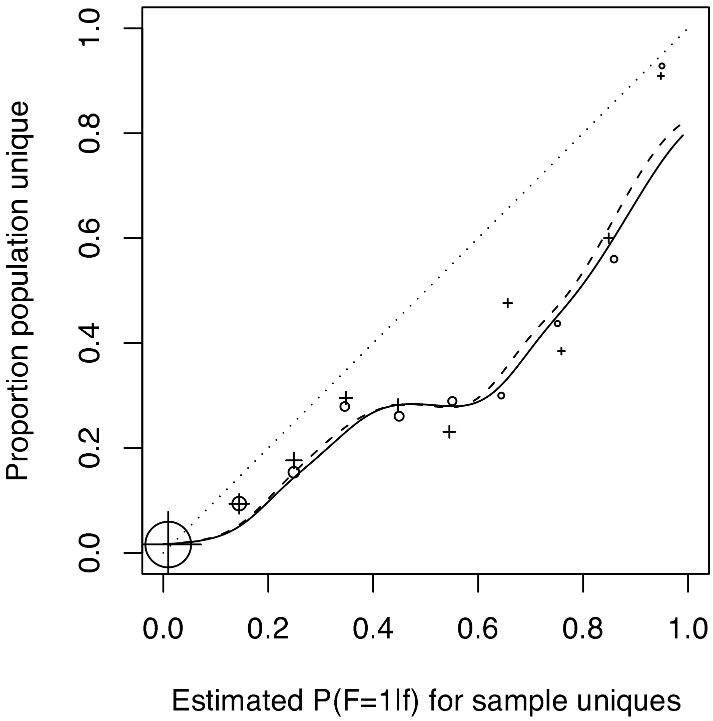

Fig. 4. The proportion of population uniques plotted against the average value of $P(F_i = 1|\boldsymbol{f})$, for cells categorised into 10 intervals according to value of $P(F_i = 1|\boldsymbol{f})$. The size of the plotting point is proportional to the number of cells in each of the intervals. The solid and dashed lines are kernel smooth fits through the points $(P(F_i = 1|\boldsymbol{f}), I[F_i = 1])$ for the sample uniques. The dotted line represents perfectly calibrated prediction. Three different priors are used, $p = 1$ (solid line, points plotted as o), $p = 50$ (dashed line, points plotted as +) and $p = 1/50$ (identical to $p = 1$).

overestimation, particularly for low-to-moderate risk cells is apparent when we fit a smooth curve through the points of Figure 3. Figure 3 corresponds to the prior with $p = 1$. Very similar results were obtained for the other two values of p.

We investigated how accurately the model can predict population unique cells. For each of the 1543 sample uniques, we calculate the predictive probability of population uniqueness, $P(F_i = 1|\boldsymbol{f})$. Figure 4 summarises the accuracy of the prediction. We divided the probability range into ten equal intervals, and for sample unique cells with $P(F_i = 1|\boldsymbol{f})$ in each of these intervals, we plotted the proportion of population uniques against the average value of $P(F_i = 1|\boldsymbol{f})$. We also plotted kernel smooth fits through the points $(P(F_i = 1|\boldsymbol{f}), I[F_i = 1])$ for the sample uniques. For accurate estimation we would expect the points and the smooth fit to lie close to the dotted line on Figure 4; the plot again reveals a tendency to slightly overestimate the probability of population uniqueness, most evident in the cells with risk values in the middle of the range. As with the overall measures of disclosure risk, there is negligible difference between $p = 1/50$ and $p = 1$, and more obvious sensitivity to higher values of p.

The analysis above is based on a single example. To more fully evaluate the performance,

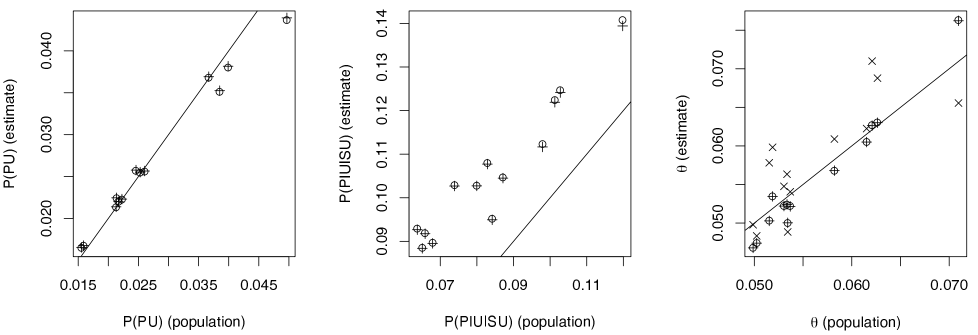

Fig. 5. The estimated overall measures of disclosure risk, $P(PU)$, $P(PU|SU)$ and θ, plotted against the population values for the 13 UK regions. The Monte Carlo estimates are plotted as +, those based on the approximations as o. The line on each plot represents equality (no error). For θ, we also plot (as x) the estimate of Skinner and Elliot (2002), presented in (4).

both of the approximations we have proposed, and of the overall methodology for disclosure risk assessment, we carried out a more extensive study using all 13 UK regions. As above, for each region the SAR data were considered as the population and a 3% sample extracted for disclosure analysis. Henceforth, we focus on a single prior using our default value $p = 1$.

First, we considered the overall disclosure risk measures $P(PU)$, $P(PU|SU)$ and θ. For each of these three measures, Figure 5 presents a plot of the estimate against the population value, for both Monte Carlo and approximation-based estimates. The first thing to notice is that the Monte Carlo and approximation-based estimates are virtually identical in all cases, so the approximation can be used with a high level of confidence. (For record-level measures, the approximations were also strongly validated, with all UK regions producing plots similar to Figure 1). For $P(PU)$ and θ, the methodology also performs well at providing accurate estimates of the population quantities, with errors being small relative to the variability of the measures between regions. For θ, the model-based estimates are generally more accurate than the (admittedly more straightforward to evaluate) estimator of Skinner and Elliot (2002), presented in (4). The methodology performs least well for $P(PU|SU)$ where the risk is consistently overestimated, although the ranking of regions in terms of this measure is quite accurate. Nevertheless, we consider θ to be the most relevant and practically useful measure of overall disclosure risk, and are reassured by the generally accurate estimation of this quantity.

As an assessment of record-level disclosure risk estimation, we evaluated how well our computed $P(F_i = 1|\boldsymbol{f})$ performed as a classifier of risky cells in terms of its sensitivity and specificity. To summarise this, for each of our 13 regions, we plotted the ROC curve for classification of population uniques by $P(F_i = 1|\boldsymbol{f})$. These curves are presented in Figure 6. Although far from perfect, it indicates that $P(F_i = 1|\boldsymbol{f})$ is of considerable value as a risk classifier. For all regions except one, there exists a value (approximately 0.12) at which over 80% of population uniques are correctly identified, with under 20% false positives. The outlying curve in Figure 6 is for Northern Ireland (the smallest region) where prediction of uniques is slightly less good.

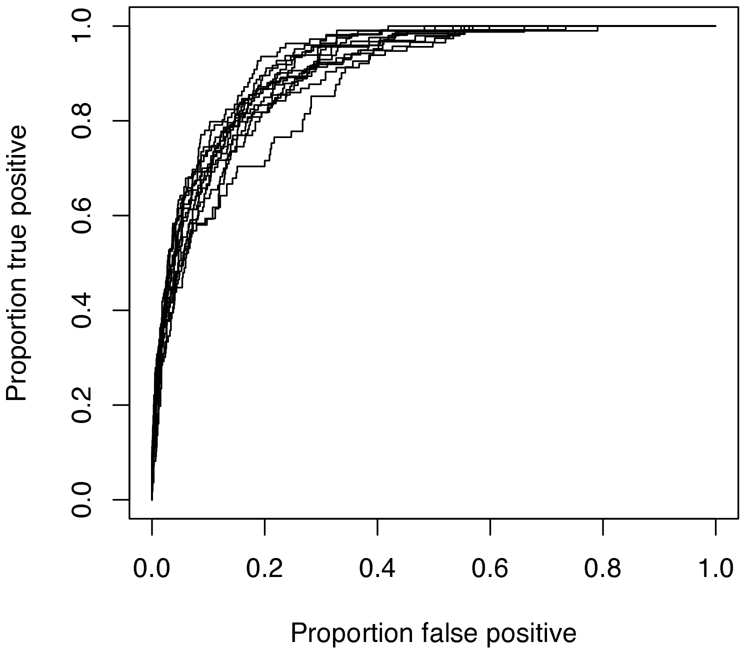

Fig. 6. ROC curves for classifying population uniques according to $P(F_i = 1|\boldsymbol{f})$. The curves illustrate how the proportions of true positive identifications of population uniques and false positive identifications vary, as the threshold value of $P(F_i = 1|\boldsymbol{f})$ for classifying a cell as population unique is increased. Curves for all 13 UK regions are superimposed on the same axes.

6. Discussion

Disclosure risk estimation is a hard statistical problem, in that we are forced to make inferences where we necessarily have little information with which to do so. Hence, it is essential that we borrow strength so that, as much as possible, we learn about areas of tables with sparse data using areas with greater proportions of the population. The approach used in this paper provides a framework with which to do this. The use of models is an essential component, and we use the sample data to identify good areas of model space, averaging across models where uncertainty exists. Bayesian predictive inference provides a natural way of thinking about disclosure risk measures as predictive probabilities of disclosure events, and such probabilities can be computed naturally within a framework involving model uncertainty.

In the paper, we have used approximations to speed up our calculations, avoiding the use of Monte Carlo wherever possible. The motivation for this is to make the methodology as efficient as possible for use in scenarios where large numbers of disclosure risk calculations need to be carried out on a routine basis. All computations were performed in R (see `http://www.r-project.org/`) and took, at the most, a few seconds to complete.

The examples presented in Section 5 illustrate that this approach has potential for identifying cells which may pose a disclosure risk. There is some evidence that the approach can be slightly conservative, tending to overestimate the risk posed by sensitive cells. It is possible that this is due to oversmoothing in the estimation procedure. The sample zero cells, of which there are a large number, will all have posterior means for F_i greater than zero. The greater the degree of this smoothing, the correspondingly smaller will be the posterior means for cells with positive sample cell counts, including the risky cells. Hence, if the degree of shrinkage implicit in the choice of prior distribution is too severe, then this may lead to an overestimation of risk. The proposal in this paper is to use a symmetric Dirichlet prior with weight equivalent to that of a single observation, for which the resulting estimates are largely satisfactory. Nevertheless, this remains an area of further research.

Acknowledgement

This work is based on the SARs provided through the Centre for Census and Survey Research of the University of Manchester with the support of ESRC and JISC. We are grateful to Julian Stander and two referees, for helpful comments on earlier versions of this paper.

References

Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Functions.* Dover, New York.

Benedetti, R and Franconi, L. (1998). Statistical and technical solutions for controlled data dissemination. In *Pre-Proceedings of New Techniques and Technologies for Statistics, Volume 1* 225–232. Sorrento, Italy.

Bethlehem, J. G., Keller, W. J. and Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, **85**, 38–45.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L. and Spiegelhalter, D. J. (1996). *Probabilistic Networks and Expert Systems.* Springer-Verlag, New York.

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, **21**, 1272–1317.

Elamir, E. A. H. and Skinner, C. J. (2004). Record-level measures of disclosure risk for survey microdata. *S^3RI Methodology Working Paper*, **M04/02**. Southampton Statistical Sciences Research Institute.

Ericson , W. A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society, B*, **31**, 195–224.

Fan, D. Y. (1991). The distribution of the product of independent beta variables. *Communications in Statistics – Theory and Methods*, **20**, 4043–4052.

Fienberg, S. E. and Makov, U. E. (1998). Confidentiality, uniqueness and disclosure limitation for categorical data. *Journal of Official Statistics*, **14**, 385–397.

Hankin, R. K. S. (2005). *The Davies Package.*, available at
`http://cran.r-project.org/doc/packages/Davies.pdf`

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, **14**, 382–401.

Madigan, D. and Raftery, A. E. (1995). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, **89**, 1535–1546.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review* **63**, 215–232.

Office for National Statistics and University of Manchester. (2005). *2001 Individual Sample of Anonymised Records (Licensed File) [computer file].* Office for National Statistics, Census Division, [original data producer(s)]. University of Manchester, Cathie Marsh Centre for Census and Survey Research [distributor].

Omori, Y. (1999). Measuring identification disclosure risk for categorical microdata by posterior population uniqueness. In *Proceedings of the International Conference on Statistical Data Protection SDP '98*, 59–76. Eurostat, Luxembourg.

Polettini, S. (2003). Some remarks on the individual risk methodology. In *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality.* Luxembourg.

Polettini, S. and Stander, J. (2004). A hierarchical Bayesian model approach to risk estimation in statistical disclosure limitation. In *Privacy in Statistical Databases*, J Domingo-Ferrer and V Torra (Eds), 247–261. Springer Lecture Notes in Computer Science, 3050, Berlin.

Rinott, Y. (2003). On models for statistical disclosure risk estimation. In *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality.* Luxembourg.

Skinner, C. J. and Elliot, M. J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society, B*, **64**, 855–867.

Skinner, C. J. and Holmes, D. J. (1998). Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, **14**, 361–372.

Takemura, A. (1999). Some superpopulation models for estimating the number of population uniques. In *Proceedings of the International Conference on Statistical Data Protection SDP '98*, 45–58. Eurostat, Luxembourg.