STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:3443–3458
Published online 11 January 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2815

Analysis on binary responses with ordered covariates and missing data

Jeremy M. G. Taylor[1,*,†], Lu Wang[2] and Zhiguo Li[1]

[1]*Department of Biostatistics, 1420 Washington Heights, University of Michigan, Ann Arbor, MI 48109, U.S.A.*
[2]*Department of Biostatistics, Harvard University, Boston, MA 02115, U.S.A.*

SUMMARY

We consider the situation of two ordered categorical variables and a binary outcome variable, where one or both of the categorical variables may have missing values. The goal is to estimate the probability of response of the outcome variable for each cell of the contingency table of categorical variables while incorporating the fact that the categorical variables are ordered. The probability of response is assumed to change monotonically as each of the categorical variables changes level. A probability model is used in which the response is binomial with parameters p_{ij} for each cell (i, j) and the number of observations in each cell is multinomial. Estimation approaches that incorporate Gibbs sampling with order restrictions on p_{ij} induced *via* a prior distribution, two-dimensional isotonic regression and multiple imputation to handle missing values are considered. The methods are compared in a simulation study. Using a fully Bayesian approach with a strong prior distribution to induce ordering can lead to large gains in efficiency, but can also induce bias. Utilizing isotonic regression can lead to modest gains in efficiency, while minimizing bias and guaranteeing that the order constraints are satisfied. A hybrid of isotonic regression and Gibbs sampling appears to work well across a variety of scenarios. The methods are applied to a pancreatic cancer case–control study with two biomarkers. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: isotonic regression; order restrictions; biomarkers; parameter constraints; multiple imputation

1. INTRODUCTION

In many situations in biomedical studies, the population can be grouped according to some ordered categorical covariates. In some applications it will be natural or reasonable to assume the mean of an outcome in the subgroups is ordered with regard to the categorical covariates.

*Correspondence to: Jeremy M. G. Taylor, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029, U.S.A.
†E-mail: jmgt@umich.edu

Contract/grant sponsor: National Cancer Institute; contract/grant number: CA97248

Received 27 September 2005
Accepted 20 November 2006

The particular research area that motivated this work comes from studies involving cancer biomarkers. There is considerable interest in discovering and assessing the molecular properties of tumors, normal tissues and serum from cancer patients and relating these properties to outcome variables, such as response to treatment or survival or case–control status. The researcher will frequently store specimens, such as a piece of the tumor or normal tissue after surgery, or a vial of serum for each patient. These specimens are later tested to determine specific molecular properties. The particular application we discuss later is from a case–control study of pancreatic cancer, with two serum biomarkers measured. The two biomarkers are CA-19-9 and CA-125, which are known to be relevant in the development and progression of pancreatic cancer. It is biologically reasonable to assume that the probability of being a case changes monotonically as the biomarker values change. Furthermore, since these two biomarkers measure different aspects of the biology of cancer, it is plausible that a combination of them may be useful for predicting the outcome variable. The overall goal is to understand the relationship between the outcome variable and the combination of covariates while utilizing the fact that the covariates are ordered. By utilizing the ordering we hope to be able to gain efficiency, compared to ignoring the ordering; this may be particularly useful in small studies.

In studies of this type missing data in one or both of the biomarkers is common. Sometimes the assay does not work for biological reasons, sometimes the specimen is missing, or degraded too much or of insufficient volume for the assay to run. Since the response is measured and one of the biomarkers may be measured it would be inappropriate to discard the observation.

There is a considerable statistical literature on statistical models and methods for ordered categorical variables and inference in the presence of monotonicity or order restrictions [1–8]. In this paper we will focus on the situation of a response variable Y and one or more ordered categorical explanatory variables X, and the general monotonicity constraint we are interested in is that if $x_1 \leqslant x_2$ then $E(Y|X=x_1) \leqslant E(Y|X=x_2)$.

Isotonic regression is a well-known approach for estimation in a regression model with a single explanatory variable and a continuous response. The pooled adjacent violators algorithm ensures that the response function is a monotonic function of the explanatory variable. The asymptotic convergence of the estimator does not follow the usual root n rate, this presents a problem for calculating standard errors and confidence intervals, particularly in small samples. If there are two or more explanatory variables the concept of isotonic regression generalizes quite naturally, although the algorithms to estimate the response surface are considerably more complex [9].

In a Bayesian approach, in general the ordering can be introduced through prior distributions. For example, if the order restriction is on the parameters of the model, say $\theta_1 < \theta_2$, then an appropriate prior would have $P(\theta_1 < \theta_2) = 1$. If it is possible to obtain draws of θ_1 and θ_2 from the posterior distribution without the order restriction that $\theta_1 < \theta_2$, then it is a simple matter to discard draws that violate the restriction to obtain draws from the desired posterior distribution. For example, in the Gibbs sampling scheme, the parameter θ_1 is drawn from its unconstrained conditional posterior distribution, but then is discarded if it is greater than the current value of θ_2 and a new value of θ_1 is drawn until one satisfying the constraint $\theta_1 < \theta_2$ is obtained [4]. This is followed by a draw of θ_2 which must be larger than the latest value of θ_1, and so on.

In a recent article Dunson and Neelon [10] developed a hybrid of isotonic regression and Gibbs sampling. In particular, they fit a model without order restrictions using Bayesian methods, but also apply isotonic regression within the Gibbs sampling algorithm. We will consider an adaptation of this as one of our approaches.

DOI: 10.1002/sim

There has been a substantial amount of research into methods for analysing data with missing values [11]. General approaches if the missingness is ignorable are through model-based schemes using maximum likelihood or Bayes estimation, through multiple imputation and through inverse probability weighting. Model-based methods can be used if the quantity of interest is a specific parameter of the parametric model. Multiple imputation is a two-stage procedure in which the missing values are first filled in, and then the augmented data is analysed. The values to be imputed are usually based on a model, called the imputing model, which may differ from the model which is used for analysis of the augmented data. An important issue in multiple imputation is choosing a reasonable imputing model. This model can be fit to the observed data using maximum likelihood or Bayesian methods. If Gibbs sampling is used to fit this model and the missing values are treated as parameters, then draws of these missing values can be used as the imputes in the analysis stage.

In this paper we consider the specific situation of a binary outcome variable and two ordered categorical covariates that could contain missing values, and develop a number of different approaches. The primary parameters of interest are the probability of response for each cell of the contingency table of categorical covariates. In Sections 2 and 3, the models and methods are developed and described. In Section 4, we present results of simulation studies comparing the methods. In Section 5 we apply the methods to the pancreatric cancer case–control study. In Section 6 we provide a summary discussion.

2. MODELS AND NOTATION

Denote the outcome as Y and the two categorical covariates as R and C, where $Y=0$ or 1, $R=1,\ldots,r$ and $C=1,\ldots,c$, and define

$$p_{ij}=P(Y=1|R=i,C=j)$$

Conditional on the total number of observations, assume a multinomial distribution for the joint distribution of R and C, with

$$q_{ij}=P(R=i,C=j)$$

where $\sum_{ij}q_{ij}=1$. Let $P=\{p_{ij}\}$ and $Q=\{q_{ij}\}$. Observations can be grouped into a two-way contingency table labelled by the row variable R and the column variable C. However, the value of R, C or both may be missing for some observations. Thus we observe either Y, R and C, or Y and R, or Y and C, or just Y. Since in many situations, the missingness is not related to the value of R and C, we can assume missing is random. The aim is to estimate p_{ij}, with q_{ij} being regarded as nuisance parameters that are necessary to include when there is missing data. We assume that the response probability will increase as R increases and also increase as C increases, that is, a partial order exists across the cells both vertically and horizontally. In particular we assume $p_{i,j}\leqslant p_{i,j+1}$ and $p_{i,j}\leqslant p_{i+1,j}$ for all i and j.

Let n_{ij} be the number of observations and d_{ij} be the number of responses in cell ij. Let $\widehat{p}_{ij}$ denote a point estimate of p_{ij} and $\widehat{\mathrm{se}}_{ij}$ denote an estimated standard error of $\widehat{p}_{ij}$.

When the missingness mechanism is missing completely at random, the complete-case analysis which uses only observations in which Y, R and C are all measured provides valid results, although it is not efficient if the fraction of complete cases is small. There is information about p_{ij} in the partially missing observations, which together with the order restrictions has the potential to enhance the efficiency of the method.

DOI: 10.1002/sim

3. ESTIMATION METHODS

For comparison purposes we will be describing methods that do not guarantee ordering amongst the estimates of p_{ij}'s as well as those that do. For each of the methods we describe it first in the case of no missing data, and then how it can be adapted to handle missing data. We will assume missingness is missing at random.

In the presence of missing covariate data we consider two strategies. One is model-based in which the algorithms are extended to allow for missing data and inference is based on the parameter estimates derived from fitting the model. The second strategy is based on multiple imputation. In this approach the observations with missing covariate values are imputed into cells of the contingency table, then the augmented data is analysed using approaches appropriate for no missing data situations. The multiple imputation approach can incorporate the order restrictions either in the model of the imputation step or in the analysis step or both.

3.1. Bayesian model-based approach

The order restrictions on the parameters are induced using the prior distributions. Both strong and weak priors are considered, for a strong prior the parameter space has positive probability only on regions that satisfy the constraints, for the weak prior the ordering is incorporated by making the prior distributions stochastically ordered.

3.1.1. No missing data. We consider three different priors for p_{ij}

- *No ordering*: Assume p_{ij} are iid $\sim$Beta(1, 1).
- *Weak ordering*: Assume p_{ij} are iid $\sim$Beta($\alpha_{ij}, 2-\alpha_{ij}$), where α_{ij} are known constants satisfying the partial order restriction. We note that $\alpha_{ij}/2$ is the mean of the prior distribution, thus the prior distributions are stochastically ordered.
- *Strong ordering*: Assume the set of p_{ij}'s satisfy the partial order restriction, that is the prior for the set of p_{ij}'s is proportional to $I(P_{r\times c} \in \text{CS}) \prod \text{Beta}(\alpha_{ij}, 2-\alpha_{ij})$, where CS is the constrained space and I denotes the indicator function. Here we might take α_{ij} as all equal or having an ordering themselves.

The prior for q_{ij} is: $\{q_{ij}\}\sim$Dirichlet($\gamma_{11}, \gamma_{12} \ldots \gamma_{1c} \ldots \gamma_{r1}, \gamma_{r2} \ldots \gamma_{rc}$), where $\{\gamma_{ij}\}$ are known constants. We note that with no missing data the prior for q_{ij} is not needed because the posterior distributions for p_{ij} and q_{ij} are independent, but the posterior distributions will not be independent when there are missing covariates.

These different prior distributions for p_{ij} will induce different properties in terms of the efficiency and bias. Only the strong ordering prior will guarantee an ordering of the estimates. The weak ordering prior will encourage, but not force the estimates of p_{ij} to be correctly ordered. In large samples we would expect the data to dominate the prior, and thus given that the true underlying probability distribution does satisfy the order restriction, we would expect all the estimates to satisfy the order constraint.

Gibbs sampling is used to estimate the parameters. It consists of drawing each parameter p_{ij} and q_{ij} from the conditional posterior distribution given the data and all the other parameters.

For p_{ij}, the conditional posterior distribution will be different for the three priors.

DOI: 10.1002/sim

For weak ordering prior

$$p_{ij}|(\text{all other parameters, data}) \sim \text{Beta}(\alpha_{ij} + d_{ij}, 2 - \alpha_{ij} + n_{ij} - d_{ij}) \tag{1}$$

where α_{ij} are known. The no ordering prior is a special case of this with $\alpha_{ij} = 1$.

For strong ordering prior

$$p_{ij}|(\text{all other parameters, data}) \sim \text{Beta}(\alpha_{ij} + d_{ij}, 2 - \alpha_{ij} + n_{ij} - d_{ij}) \tag{2}$$

and

$$\max(p_{i-1,j}^{(k)}, p_{i,j-1}^{(k)}) < p_{ij} < \min(p_{i+1,j}^{(k-1)}, p_{i,j+1}^{(k-1)}) \tag{3}$$

where $p_{i-1,j}^{(k)}$ and $p_{i,j-1}^{(k)}$ are draws from the current iteration and $p_{i+1,j}^{(k-1)}$ and $p_{i,j+1}^{(k-1)}$ are Gibbs draws from the previous iteration. That is, we draw a value of p_{ij} from a truncated beta distribution, where the limits of the distribution are defined by the order restrictions with the current values of the parameters.

It is easy to see that

$$Q|P, \text{data} \sim \text{Dirichlet}(\gamma_{11} + n_{11}, \ldots, \gamma_{rc} + n_{rc}) \tag{4}$$

For all the Gibbs sampling schemes we use the mean of the posterior draws as the point estimate and the range between the 2.5 per cent quantile and the 97.5 per cent quantile as a 95 per cent interval for the parameters. We typically draw 2500 samples and discard the first 500.

3.1.2. Missing data. The parameters of the model can be estimated by using a data augmentation algorithm [11]. Regarding the missing data also as parameters, let $X_{\text{obs}} = \{R_{\text{obs}}, C_{\text{obs}}, Y\}$, $X_{\text{mis}} = \{R_{\text{mis}}, C_{\text{mis}}\}$, $\theta = \{P = \{p_{ij}\}, Q = \{q_{ij}\}, i = 1, \ldots, r, j = 1, \ldots, c\}$. We extend the Gibbs sampling algorithm in 3.1.1 such that in the kth iteration, we conduct the following I and P steps. These steps are repeated until convergence.

1. *I-step*: For each observation with missing values we impute the values of R and C from the following multinomial distributions.
   For the observations with known R and unknown C

$$P(C = j|P^{(k-1)}, Q^{(k-1)}, Y = y, R = i) = \frac{(1 - p_{ij}^{(k-1)})^{1-y} * (p_{ij}^{(k-1)})^{y} * q_{ij}^{(k-1)}}{\sum_{l=1}^{c} (1 - p_{il}^{(k-1)})^{1-y} * (p_{il}^{(k-1)})^{y} * q_{il}^{(k-1)}}$$

   with an analagous expression for when C is known and R unknown.
   For the observations with unknown R and unknown C

$$P(R = i, C = j|P^{(k-1)}, Q^{(k-1)}, Y = y) = \frac{(1 - p_{ij}^{(k-1)})^{1-y} * (p_{ij}^{(k-1)})^{y} * q_{ij}^{(k-1)}}{\sum_{m=1}^{r} \sum_{n=1}^{c} (1 - p_{mn}^{(k-1)})^{1-y} * (p_{mn}^{(k-1)})^{y} * q_{mn}^{(k-1)}}$$

   where $i = 1, \ldots, r$ and $j = 1, \ldots, c$.
2. *P-step*: Generate $\theta^{(k)}$ from $P(\theta|X_{\text{obs}}, X_{\text{mis}}^{(k)})$, where $X_{\text{mis}}^{(k)}$ is the value obtained in the I-step. This process is implemented by using an adapation of the Gibbs sampling algorithm as described earlier. Specifically, we use $n_{ij}^{(k)}$ to denote the total number of observations in the cell with $R = i$, $C = j$ and use $d_{ij}^{(k)}$ to denote the number of

DOI: 10.1002/sim

responses in that cell after using draws from the fully conditional distribution to impute the data in the I-step. Then the conditional posterior distribution for $Q|(P, X_{\text{obs}}, X_{\text{mis}}^{(k)})$ and $p_{ij}|$(all other parameters, $X_{\text{obs}}, X_{\text{mis}}^{(k)}$) are the same as given in equations (1)–(4) except that $d_{ij}^{(k)}$ replaces d_{ij} and $n_{ij}^{(k)}$ replaces n_{ij}.

3.2. Empirical estimator

3.2.1. No missing data. The simplest estimator is $\widehat{p}_{ij} = d_{ij}/n_{ij}$ with $\widehat{\text{se}}_{ij} = \sqrt{(\widehat{p}_{ij}(1-\widehat{p}_{ij})/n_{ij})}$. We note that this estimator will not necessarily give estimates of p_{ij} satisfying the order restriction.

3.2.2. Missing data. With missing data the proposed method is to apply the empirical estimator to multiply imputed data sets. The missing values are imputed K times using the model-based Gibbs sampling algorithm as described in Section 3.1.2.

The estimates and standard errors from the kth completed data set are given by $\widehat{p}_{ij}^{(k)} = d_{ij}^{(k)}/n_{ij}^{(k)}$ and

$$\widehat{\text{se}}_{ij}^{(k)} = \sqrt{\frac{(d_{ij}^{(k)} + \alpha_{ij}) \times (n_{ij}^{(k)} - d_{ij}^{(k)} + \beta_{ij})}{(n_{ij}^{(k)} + \alpha_{ij} + \beta_{ij})^2 \times (n_{ij}^{(k)})}}$$

Then results from the K data sets are combined according to the following standard rules for multiple imputation.

The final estimate of p_{ij} is $\widehat{p}_{ij} = \sum_{k=1}^{K} \widehat{p}_{ij}^{(k)}$, and the estimate of the variance of $\widehat{p}_{ij}$ is $V_{ij} = W_{ij} + ((K+1)/K)B_{ij}$, where the average with-imputation variance is $W_{ij} = \frac{1}{K}\sum_{k=1}^{K}(\widehat{\text{se}}_{ij}^{(k)})^2$, and the between-imputation variability is $B_{ij} = 1/(K-1)\sum_{k=1}^{K}(\widehat{p}_{ij}^{(k)} - \widehat{p}_{ij})^2$. A 95 per cent interval is calculated as $CI_{ij} = \widehat{p}_{ij} \pm 1.96 \times \sqrt{V_{ij}}$.

3.3. Modified estimator

3.3.1. No missing data. In small samples the simple empirical estimator can have a standard error of zero if $d_{ij} = 0$ or n_{ij}. Also a typical confidence interval based on + or − 1.96 standard errors frequently goes outside the range 0 to 1. To overcome some of these problems a slightly modified version of the simple empirical estimator is an alternative. Motivated by a Bayesian approach with a Beta(α_{ij}, β_{ij}) prior distribution, the estimator and standard error are

$$\widehat{p}_{ij} = \frac{d_{ij} + \alpha_{ij}}{n_{ij} + \alpha_{ij} + \beta_{ij}}$$

and

$$\widehat{\text{se}}_{ij} = \sqrt{\frac{(d_{ij} + \alpha_{ij}) \times (n_{ij} - d_{ij} + \beta_{ij})}{(n_{ij} + \alpha_{ij} + \beta_{ij})^2 \times (n_{ij} + \alpha_{ij} + \beta_{ij} + 1)}} = \sqrt{(\widehat{p}_{ij}(1-\widehat{p}_{ij})/(n_{ij} + \alpha_{ij} + \beta_{ij} + 1))}$$

We will consider two cases, one where $\alpha_{ij} = \beta_{ij} = 1$, and one where $\beta_{ij} = 2 - \alpha_{ij}$ and the α_{ij} are ordered. In this latter case the mean of the prior distribution is $\alpha_{ij}/2$, and these will be selected to match prior beliefs about the values of p_{ij}.

DOI: 10.1002/sim

3.3.2. Missing data. Multiple imputation is used to impute the missing values, followed by application of the modified estimator as described above.

After convergence of the Gibbs sampler at each iteration we estimate p_{ij} using $\widehat{p}_{ij}^{(k)} = (d_{ij}^{(k)} + \alpha_{ij}) / (n_{ij}^{(k)} + \alpha_{ij} + \beta_{ij})$ and estimate se_{ij} using

$$\widehat{\mathrm{se}}_{ij}^{(k)} = \sqrt{\frac{(d_{ij}^{(k)} + \alpha_{ij}) \times (n_{ij}^{(k)} - d_{ij}^{(k)} + \beta_{ij})}{(n_{ij}^{(k)} + \alpha_{ij} + \beta_{ij})^2 \times (n_{ij}^{(k)} + \alpha_{ij} + \beta_{ij} + 1)}}$$

These estimates are then combined using the standard rules for combining estimates from multiply imputed data sets.

3.4. Isotonized estimators

3.4.1. No missing data. An adaptation of the empirical or modified estimator is to follow it by isotonic regression. Isotonic regression in greater than one dimension can be described in the following general way. For $X = \{x_1, x_2, \ldots, x_k\}$, w is a positive weight function defined on X, F is a restricted family of functions on X for arbitrary function g defined on X. A function g^* on X is an isotonic regression of g with weights w if and only if g^* is isotonic and g^* minimizes $\sum_{x \in X} [g(x) - f(x)]^2 w(x)$ in the class of all isotonic functions $f \in F$. We implement isotonic regression in two dimensions using the algorithm described in [9]. The final estimate is denoted by $\widehat{p_{ij}}^* = g^*(\widehat{p_{ij}})$, where g^* is the isotonic regression transformation.

We note that while this method guarantees the order restriction is satisfied, there is no simple way to obtain its standard error. We use $\sqrt{(\widehat{p}_{ij}^*(1 - \widehat{p}_{ij}^*)/n_{ij})}$ or $\sqrt{(\widehat{p}_{ij}^*(1 - \widehat{p}_{ij}^*)/(n_{ij} + \alpha_{ij} + \beta_{ij} + 1))}$ as the standard error and evaluate this approximation in a simulation study.

3.4.2. Missing data. Multiple imputation is used to impute the missing values, followed by application of the isotonized estimator as described above.

After convergence of the Gibbs sampler, at each iteration we obtain an estimate $\widehat{p}_{ij}^{(k)}$ of p_{ij} and this is then isotonized using the transformation $\widehat{p}_{ij}^{(k)*} = g^*(\widehat{p}_{ij}^{(k)})$.

3.5. Isotonized Gibbs sampling

3.5.1. No missing data. Of the Bayesian model-based approaches only the strong ordering Gibbs sampling approach guarantees that the order constraints are satisfied. One way to force the correct ordering of point estimates in cases where it is not guaranteed is to incorporate the isotonic regression transformation strategy into the Gibbs sampling iterations. Specifically, at the kth iteration, denote the set of draws from the unconstrained posterior distributions as $\{p_{ij}^{(k)}\}$. Let $p_{ij}^{(k)*} = g^*(p_{ij}^{(k)})$, where g^* is the isotonic regression transformation. Thus, we project draws from a unconstrained posterior distribution onto an order-restricted parameter space using a minimal distance mapping. Then we focus on $p_{ij}^{(k)*}$ to make statistical inference. This is an extension of the method described by Dunson and Neelon [10]. We use the sample mean of $p_{ij}^{(k)*}$ as the final estimate of the parameters and use the range between the 2.5 per cent quantile and the 97.5 per cent quantile of $p_{ij}^{(k)*}$ as a 95 per cent interval for the parameters. It is not clear whether this hybrid approach corresponds to a specific statistical model, nevertheless it is a method that can

DOI: 10.1002/sim

be applied and evaluated in a simulation study for example. It is somewhat similar to an EMS strategy, where a smoothing step is introduced within the EM algorithm [12].

3.5.2. Missing data. In the case of missing data we use the Gibbs sampling appropriate for missing data, as described in Section 3.1.2, with the additional isotonic step as described above. This approach gives values of $p_{ij}^{(k)*}$ which are then treated as if they were draws from a posterior distribution.

4. SIMULATION STUDIES

4.1. Data generation

To compare the above methods, we conducted simulations using 3×2 tables ($r = 3$, $c = 2$). We generated 200 data sets according to different sets of true values $\{p_{ij}^0\}$. The values of $\{q_{ij}^0\}$ are set to $\frac{1}{6}$ in all cases. Each table contains either 60, 150 or 300 observations. We set the missing proportion as either 0, 20 or 40 per cent with a missingness mechanism as missing completely at random.

4.2. Methods compared

For no missing data situations the methods considered were empirical, isotonized empirical, modified, isotonized modified, as described in Sections 3.2.1, 3.3.1, 3.4.1, and Gibbs(no), Gibbs(weak) with $\alpha_{ij} = 2p_{ij}^0$ and Gibbs(strong) with $\alpha_{ij} = 2p_{ij}^0$, as described in Section 3.1.1, and Isotonized-Gibbs(no) and IsotonizedGibbs(weak), where IsotonizedGibbs is Gibbs sampling with isotonic transformation at each iteration as described in Section 3.5.1.

For missing data situations the model-based methods considered were Gibbs(no), Gibbs(weak), Gibbs(strong), IsotonizedGibbs(no) and IsotonizedGibbs(weak). The multiple imputation schemes are denoted as MI(a,b), where 'a' refers to the estimator which could be empirical (em), isotonized empirical(iso_em), modified(mo) or isotonized modified (iso-mo) and 'b' refers to the prior distribution, which could be no, weak or strong. Also included for comparison purposes are the results for the empirical estimator applied to the data before any of it was made missing.

4.3. Measures of comparison

For each method and each data set s, we obtained for each cell (i, j) the point estimates $p_{ij\,(s)}$ and the corresponding standard error $\text{se}_{ij\,(s)}$, and an indicator variable $\text{cove}_{ij\,(s)}$ for whether the 95 per cent interval contained the true value. We simulated 200 data sets.

Let $\widetilde{p}_{ij} = \frac{1}{200}\sum_{s=1}^{200} p_{ij\,(s)}$ denote the average for cell (i, j) and the coverage rate is denoted by $\widetilde{\text{cove}}_{ij} = \frac{1}{200}\sum_{s=1}^{200}\text{cove}_{ij\,(s)}$. The empirical variance of the point estimates is given by $\text{var}_{ij} = \frac{1}{199}$ $\sum_{s=1}^{200}(p_{ij\,(s)} - \widetilde{p}_{ij})^2$.

We consider the following measures of overall performance of the various methods:

Average bias $= \sum_i \sum_j |\widetilde{p}_{ij} - p_{ij}^0|/(rc)$, where p_{ij}^0 is the true value of p_{ij}.
Max bias $= \max_{ij}(|\widetilde{p}_{ij} - p_{ij}^0|)$
Average coverage $= \sum_i \sum_j \widetilde{\text{cove}}_{ij}/(rc)$

DOI: 10.1002/sim

Min coverage $= \min_{ij}(\widetilde{\text{cove}}_{ij})$
Empirical efficiency $= (\sum_i \sum_j \frac{\text{var}_{ij}}{\text{var0}_{ij}}/(rc))^{-1}$, where var0_{ij} is the variance of the estimate of p_{ij} calculated from the empirical method applied to the no missing data situation.
Order proportion = the proportion of the inequality constraints that are satisfied by the final point estimates.
MSE $= \sum_i \sum_j (\text{var}_{ij} + (\widetilde{p}_{ij} - p^0_{ij})^2)$

4.4. Comparison of methods

Table I shows the results for no missing data, for the 3×2 table:

0.2	0.4
0.4	0.6
0.6	0.8

of true values of $\{p_{ij}\}$'s. The results show some variation between the methods, and at the smaller sample size the coverage rates can be less than the nominal level.

The Gibbs sampling method with the inclusion of strong inequality constraints can lead to substantial gains in efficiency compared to not including any ordering constraints, however this can be accompanied by an increase in bias and resulting problems with the coverage rate. Introducing

Table I. Simulation results.

Method	Average bias $\times 10^5$	Max. bias $\times 10^5$	Ave. coverage	Min. coverage	Efficiency	Order proportion	MSE $\times 10^5$
$n = 300$							
Empirical	329	974	0.92	0.89	1	0.98	2819
Modified	964	1476	0.93	0.90	1.09	0.98	2651
Isotonized empirical	325	957	0.95	0.93	0.96	1	2691
Isotonized modified	614	1329	0.94	0.93	1.22	1	1730
Gibbs(no)	737	1340	0.94	0.91	1.09	0.98	2634
Gibbs(weak)	969	1482	0.94	0.91	1.01	0.98	2651
Gibbs(strong)	962	1770	0.95	0.93	1.35	1	1975
Isotonized Gibbs(no)	573	1229	0.94	0.92	1.16	1	2235
Isotonized Gibbs(weak)	974	1734	0.94	0.91	1.15	1	2323
$n = 60$							
Empirical	416	726	0.91	0.88	1	0.83	13493
Modified	3044	4822	0.91	0.87	1.52	0.88	9526
Isotonized empirical	1186	1915	0.98	0.98	1.43	1	9608
Isotonized modified	3044	5234	0.94	0.87	1.96	1	7518
Gibbs(no)	2934	5146	0.96	0.95	1.52	0.83	9613
Gibbs(weak)	3032	4833	0.94	0.89	1.52	0.88	9523
Gibbs(strong)	4778	8291	0.92	0.81	3.23	1	5854
Isotonized Gibbs(no)	1372	2233	0.97	0.95	2.70	1	5127
Isotonized Gibbs(weak)	3062	6357	0.94	0.88	2.44	1	6340

Comparison of estimation methods. 0 per cent missing.

DOI: 10.1002/sim

Table II. Simulation results.

Method	Average bias $\times 10^5$	Max. bias $\times 10^5$	Ave. coverage	Min. coverage	Efficiency	Order proportion	MSE $\times 10^5$
$n = 300$							
Empirical complete data	329	974	0.92	0.89	1	0.98	2819
Gibbs(no)	762	1570	0.95	0.94	0.83	0.96	3191
Gibbs(strong)	1236	2616	0.96	0.94	1.32	1	2101
IsotonizedGibbs(no)	332	861	0.95	0.93	1.08	1	2443
MI(em,no)	508	1078	0.93	0.91	0.76	0.95	3441
MI(mo,no)	767	1581	0.94	0.92	0.83	0.96	3196
MI(is_em,no)	435	1034	0.95	0.94	0.96	1	2897
MI(is_mo,no)	687	1321	0.96	0.93	1.05	1	2726
MI(em,strong)	447	1052	0.95	0.91	0.88	0.98	2970
MI(mo,strong)	564	1207	0.96	0.94	0.96	0.98	2741
$n = 60$							
Empirical complete data	416	726	0.91	0.88	1	0.83	13493
Gibbs(no)	3457	7488	0.96	0.93	1.18	0.78	12724
Gibbs(strong)	4888	9385	0.92	0.72	3.03	1	6688
IsotonizedGibbs(no)	1646	2496	0.97	0.94	2.38	1	5903
MI(em,no)	823	2284	0.97	0.92	0.62	0.77	18761
MI(mo,no)	3472	7500	0.97	0.93	1.18	0.77	12711
MI(is_em,no)	1758	2816	0.98	0.96	1.32	1	10612
MI(is_mo,no)	2018	4643	0.98	0.94	2.08	1	7006
MI(em,strong)	894	1821	0.97	0.94	0.93	0.83	14784
MI(mo,strong)	2472	5557	0.98	0.95	1.49	0.84	9699

Comparison of estimation methods. 20 per cent missing.

the ordering in a weaker way, had negligible impact on the bias and efficiency even at small sample sizes and only a very small impact on causing the point estimates to be correctly ordered.

The isotonized methods did lead to estimates that were always correctly ordered, and tended to have small MSE's. The potential gain in efficiency is considerable particularly at small sample sizes.

Table II shows the results when there is 20 per cent missing data. Results from various types of Gibbs sampling and multiple imputation are included. We note that the Gibbs methods with strong ordered priors can give substantial gains in efficiency, but can also give considerable bias, leading to undercoverage. For the multiple imputation methods, using a strong ordered prior, does improve the proportion of estimates that are correctly ordered, but they are still far from perfect. The inclusion of isotonic steps within a method gives an improvement in efficiency. The Isotonized Gibbs(no) and MI(is_em,no) appeared to be promising estimators, based on giving the correct ordering, low bias, adequate coverage rates and gains in efficiency.

In Table III we consider results from a different 3×2 table:

0.4	0.5
0.6	0.9
0.8	0.91

DOI: 10.1002/sim

Table III. Simulation results.

Method	Average bias $\times 10^5$	Max. bias $\times 10^5$	Ave. coverage	Min. coverage	Efficiency	Order proportion	MSE $\times 10^5$
$n = 300$, 0 per cent missing							
Empirical	374	742	0.92	0.89	1	0.90	2386
Isotonized empirical	744	1016	0.96	0.94	1.22	1	2072
Gibbs(no)	1129	1959	0.94	0.93	1.09	0.90	2295
Gibbs(strong)	1642	2368	0.94	0.85	1.67	1	1691
Isotonized Gibbs(no)	1351	2901	0.94	0.91	1.45	1	1865
$n = 300$, 40 per cent missing							
Empirical complete data	374	742	0.92	0.89	1	0.90	2386
Gibbs(no)	1766	3684	0.95	0.93	0.62	0.85	4141
Gibbs(strong)	2743	5819	0.93	0.80	1.20	1	2670
Isotonized Gibbs(no)	2360	4948	0.94	0.87	1.01	1	2945
MI(is_em,no)	2135	4442	0.97	0.92	0.95	1	2856
$n = 150$, 40 per cent missing							
Empirical complete data	359	533	0.90	0.84	1	0.86	4750
Gibbs(no)	2953	5813	0.95	0.93	0.71	0.80	7342
Gibbs(strong)	3526	8223	0.90	0.58	1.69	1	3993
Isotonized Gibbs(no)	3414	6956	0.95	0.90	1.45	1	4351
MI(is_em,no)	2538	5536	0.97	0.96	1.14	1	4646
$n = 60$, 40 per cent missing							
Empirical complete data	431	826	0.80	0.59	1	0.81	11622
Gibbs(no)	6335	13436	0.96	0.94	1.03	0.72	14592
Gibbs(strong)	4991	13462	0.87	0.42	3.13	1	6596
Isotonized Gibbs(no)	6459	12083	0.95	0.87	2.70	1	7491
MI(is_em,no)	2151	4644	0.97	0.90	1.61	1	2854

Comparison of estimation methods. 40 per cent missing.

and focus mainly on methods that always give estimates in the correct order. Note that the table has two cells with very close probabilities. The IsotonizedGibbs(no) and MI(is_em,no) appear to have reasonable properties in this challenging case with 40 per cent missing. The bias associated with the Gibbs(strong) method can be large, resulting in very low coverage rates with small sample sizes. The Gibbs(strong) estimates in the cell with true value 0.9 tend to be much smaller than 0.9 and those in the cell with true probability of 0.91 tend to be higher than 0.91.

Overall there is no uniformly best method. Some of the methods do not guarantee the estimates will be correctly ordered. Amongst those that do give the correct ordering, the hybrid Gibbs sampling approach, with an isotonic step included in the algorithm appears to give reasonable results across a range of scenarios, as does the isotonized empirical, with multiple imputation if necessary.

In general one difference between the model-based results that used Gibbs sampling and the methods that used isotonic regression, is for parameters in neighbouring cells whose true values are very close to each other even though they satisfied the ordering constraint. The isotonic regression methods had no trouble making the estimates of these two adjacent parameters similar to each

DOI: 10.1002/sim

other. In contrast the Gibbs sampling method, in the way we implemented it where the parameter is drawn from a restricted range as in equation (3), tended to result in the parameters in adjacent cells being more widely separated.

5. APPLICATION TO PANCREATIC CANCER

The methods developed in this article were applied to a pancreatic cancer serum biomarker study [13]. This is a case–control study with 90 pancreatic cancer cases and 51 controls. Serum samples were assayed for two antigens, CA-19-9 and CA-125. To illustrate the methods we divided both biomarkers into an ordinal scale with three categories. The cut-off values for CA-19-9 are 11.2 and 28.6, and the cut-offs for CA-125 are 11.2 and 15.0. The data are presented in Table IV.

Note that each of the nine cells is labelled 1 to 9, this represents a possible overall ordering of the probability of being a case and will be later used to illustrate the estimates. We can see that both CA-19-9 and CA-125 are associated with the probability of being a case. From the data it is clear that the empirical estimates are not perfectly monotonically ordered either vertically or horizontally, and we also note that the cells labelled 3 and 5 have very small sample size. So the empirical estimate has a large standard error associated with it, but the estimates that utilize the monotonic structure are likely to be more precise because they can gain efficiency from the neighbouring cells. The estimates and standard errors from selected Gibbs sampling and isotonic methods are shown in Figures 1 and 2. The cell probability estimates are now correctly ordered both vertically and horizontally, for the methods that guarantee ordering, e.g. cell 3 is now intermediate to cells 1 and 4, and less than cell 5. The results show that the original labelling was probably a good guess at the correct overall ordering, with the possible exception of cells 7 and 8. There is considerable gains in efficiency, particularly for cells 3 and 5. Based on the simulation results, the preferred estimates for this application, would probably be the one labelled IsotonizedGibbs(no).

The data in Table IV has no missing values. To illustrate the methods when there is missing biomarker data we randomly deleted CA-125 and CA-19-9 values from this data set, with $P(\text{marker is missing}) = 0.1$ for controls and 0.2 for cases for each marker. The response Y is always observed so we retain all 141 observations. We note that this is not missing completely at random, but rather it is missing at random, for which the methods are still applicable. The new data, estimates and standard errors are given in Table V. The difference between the methods

Table IV. Pancreatic cancer biomarker data.

		CA125		
		Low	Med	High
CA19-9	Low	$\frac{2}{13}$ (1)	$\frac{3}{7}$ (3)	$\frac{6}{20}$ (4)
	Med	$\frac{1}{10}$ (2)	$\frac{1}{3}$ (5)	$\frac{8}{10}$ (8)
	High	$\frac{12}{16}$ (6)	$\frac{10}{13}$ (7)	$\frac{47}{49}$ (9)

The fractions are the number of cases out of the total number in each cell, with a cell label shown in parentheses.

DOI: 10.1002/sim

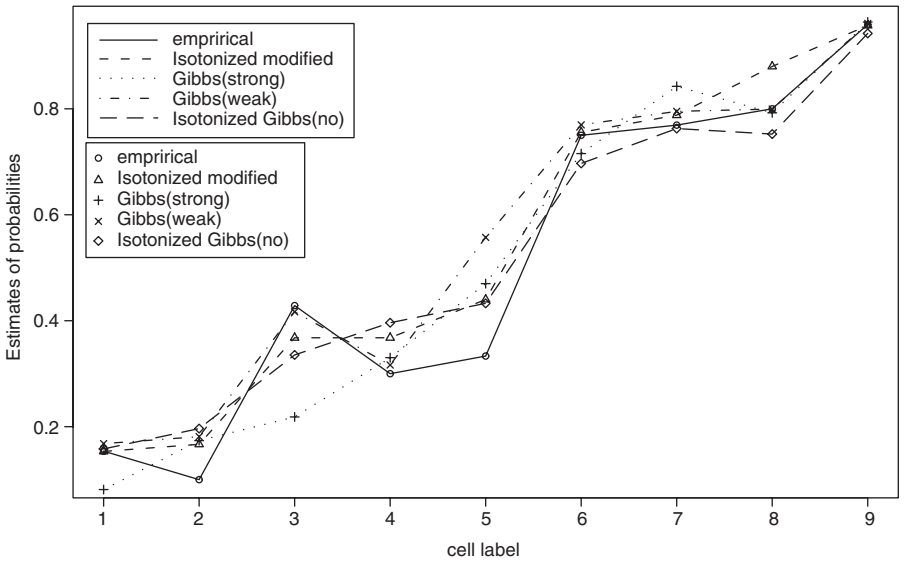

Figure 1. Estimated probabilities for nine cells.

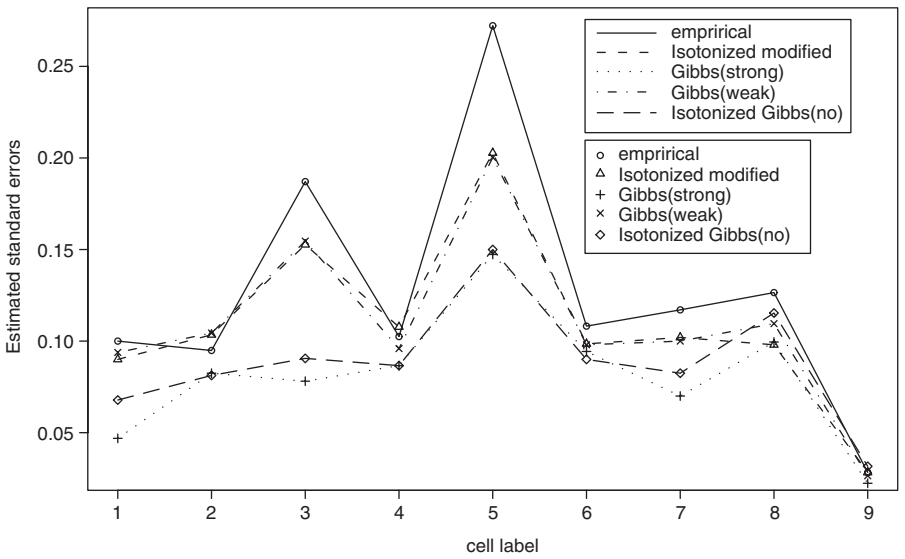

Figure 2. Estimated standard errors for nine cells.

is largest for the cells with the fewest observations, for which the point estimates between Gibbs(strong) and MI(is_em,no) can differ, and the standard errors for Gibbs(strong) tend to be smaller.

DOI: 10.1002/sim

Table V. Pancreatic cancer biomarker data.

	Cell label									
	1	2	3	4	5	6	7	8	9	Other
Data	$\frac{1}{9}$	$\frac{0}{6}$	$\frac{1}{4}$	$\frac{6}{16}$	$\frac{0}{0}$	$\frac{8}{11}$	$\frac{8}{10}$	$\frac{5}{7}$	$\frac{31}{33}$	$\frac{30}{45}$
Gibbs(strong)										
estimates	0.09	0.22	0.29	0.44	0.60	0.69	0.86	0.80	0.96	
(se)	(0.06)	(0.09)	(0.13)	(0.11)	(0.15)	(0.10)	(0.07)	(0.10)	(0.03)	
Isotonized Gibbs(no)										
estimates	0.12	0.22	0.29	0.40	0.73	0.76	0.87	0.82	0.95	
(se)	(0.07)	(0.08)	(0.12)	(0.10)	(0.18)	(0.10)	(0.06)	(0.10)	(0.03)	
MI(is_em,no)										
estimates	0.07	0.08	0.21	0.36	0.65	0.75	0.84	0.81	0.95	
(se)	(0.07)	(0.10)	(0.19)	(0.11)	(0.45)	(0.12)	(0.10)	(0.15)	(0.03)	

Results for missing data. The 'other' category denotes a combination of missing one of both of the biomarkers.

6. DISCUSSION

A different approach to data of this form is to consider Y, R and C as a rectangular array of data, with missing elements in the R and C columns. This can then be viewed as a missing data problem, for which multiple imputation would be one approach. A recently suggested method [14] consists of a sequence of regression models of each column on all the others, with the regression model used to impute the missing data. The ordered nature of R and C could be incorporated by using, for example, a cumulative proportional odds model for R or C when they are the response variable in the regression model, although this method would not guarantee the correct ordering of the cell probabilities.

There are a number of generalizations of the models and methods described here to other situations. For example, there could be more than two covariates, the outcome could be Poisson, or it could be a censored survival time. There might be other non-ordered covariates, in which case one might consider a generalized linear model such as $\text{logit}(p_{ij}) = \gamma_{ij} + \beta X$, where X are the other covariates and the order restrictions are placed on γ_{ij}.

Our simulation results indicated that no single method dominated, however methods that incorporate isotonic regression are relatively good in terms of MSE, bias, efficiency and coverage, while guaranteeing estimates in the correct order.

Among the estimate based on multiple imputation, the one that followed the imputation with an isotonic regression appeared to have good properties. In the situations we considered in the simulation there is little to recommend multiple imputation over the model-based methods, because the structure of the data is fairly simple and the models are fully saturated. However, in more complex situations with, for example, other covariates or non-completely missing at random the multiple imputation methods may be more robust.

As pointed out by Dunson and Neelon [10] one difference between the pure Bayesian approach and isotonic regression, is that isotonic regression will tend to make two parameters equal if the direction of the inequality in the order constraints is not supported by the data. In contrast, in the Bayesian scheme described by Gelfand *et al*. [4] the conditional posterior distribution for

DOI: 10.1002/sim

each parameter is restricted to a range, so it has essentially zero probability of being drawn as the limit of that range. That is, the pure Bayesian scheme is very unlikely to make $\theta_1 = \theta_2$, but this could be quite common with the isotonic regression scheme. In a recent article [15, 16] the Bayesian approach has been extended to allow equality between parameters. They proposed a prior which includes non-zero mass on the boundary. Their models were restricted to situations with one-dimensional ordering; we are currently investigating whether they can be further developed to the case of two- or higher-dimensional ordering, as one would have with multiple biomarkers.

We have restricted attention to constraints between adjacent cells, i.e. $p_{i,j} \leqslant p_{i,j+1}$ and $p_{i,j} \leqslant p_{i+1,j}$. But the methods we use could easily adapt to a set of more general constraints between parameters. In the Bayesian model-based approach these constraints could be incorporated into the prior. The methods that include an isotonizing step can also be adapted because the transformation $p_{ij}^{(k)*} = g^*(p_{ij}^{(k)})$ can be modified so that $p_{ij}^{(k)*}$ satisfies more general constraints.

The recommendation to incorporate isotonic regression into the method raises some interesting issues. One is a better justification of the method of obtaining standard errors. The way we have included isotonizing, is simply as an isotonic regression algorithm within a more standard statistical approach. We are currently investigating models which would lead more naturally to the inclusion of an isotonizing step.

A premise of this article is that monotonicity is known to exist. This may or may not be true, depending on the scientific context. However, there are many situations where it does seem *a priori* scientifically plausible. If this is the case, then incorporating it into the analysis, can lead to considerable gains in efficiency, and thus is an area worthy of further study and implementation.

ACKNOWLEDGEMENTS

This research was partially supported by grant CA97248 from the National Cancer Institute.

REFERENCES

1. Agresti A, Coull BA. Order-restricted tests for stratified comparisons of binomial proportions. *Biometrics* 1996; **52**:1103–1111.
2. Bacchetti P. Additive isotonic models. *Journal of American Statistical Association* 1989; **84**:289–294.
3. Barlow RE, Bartolomew DJ, Bremner JM, Brunk HD. *Statistical Inference under Order Restrictions*: *The Theory and Application of Isotonic Regression*. Wiley: New York, 2002.
4. Gelfand AE, Smith AFM, Lee TM. Bayesian analysis of constrained parameter and truncated data problems. *Journal of American Statistical Association* 1992; **87**:523–532.
5. Hwang JTG, Peddada SD. Confidence interval estimation subject to order restriction. *Annals of Statistics* 1994; **22**:67–93.
6. McCullagh P. Regression models for ordinal data (with Discussion). *Journal of the Royal Statistical Society*, *Series B* 1980; **42**:109–142.
7. Mukerjee H, Tu R. Order restricted inference in linear regression. *Journal of American Statistical Association* 1995; **90**:717–728.
8. Robertson T, Wright F, Dykstra R. *Order Restricted Statistical Inference*. Wiley: New York, 1988.
9. Qian S, Eddy WF. An algorithm for isotonic regression on ordered rectangular grids. *Journal of Computational and Graphical Statistics* 1996; **5**:225–235.
10. Dunson DB, Neelon B. Bayesian inference on order-restricted parameters in generalized linear models. *Biometrics* 2003; **59**:286–295.
11. Little RJA, Rubin DB. *Statistical Analysis with Missing Data* (2nd edn). Wiley: New York, 2002.
12. Silverman BW, Jones MC, Wilson JD, Nychka DW. A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography. *Journal of the Royal Statistical Society*, *Series B* 1990; **52**:271–324.

DOI: 10.1002/sim

13. Wieand S, Gail MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* 1989; **76**:585–592.
14. Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* 2001; **27**:85–95.
15. Dunson DB. Bayesian semiparametric isotonic regression for count data. *Journal of American Statistical Association* 2005; **100**:618–627.
16. Neelon B, Dunson DB. Bayesian isotonic regression and trend analysis. *Biometrics* 2004; **60**:398–406.

DOI: 10.1002/sim