Bayesian Graphical Models for Discrete Data

David Madigan[†]
University of Washington
Jeremy York[‡]
Carnegie-Mellon University

November 15, 1993

[†] Department of Statistics, GN-22, University of Washington, Seattle, WA 98195 (madigan@stat.washington.edu). Madigan's research was supported by a grant from the NSF.

[‡] York's research was supported by a NSF graduate fellowship. The authors are grateful to Julian Besag, David Bradshaw, Jeff Bradshaw, James Carlsen, David Draper, Ivar Heuch, Robert Kass, Augustine Kong, Steffen Lauritzen, Adrian Raftery, and James Zidek for helpful comments and discussions.

Contents

1 Introduction

The use of graphs to represent statistical models dates back to Wright (1921) and has been the focus of considerable activity in recent years. In particular, attention has been directed at graphical "conditional independence" models and at the application of such graphical models to probabilistic expert systems. These developments are conveniently summarised in the recent books by Whittaker (1990), Pearl (1988), and Neapolitan (1990), and in Spiegelhalter *et al.* (1993). Rather less well known are the breakthroughs that have also taken place in the development of a Bayesian framework for such models (Dawid and Lauritzen, 1993, Spiegelhalter and Lauritzen, 1990). The motivational applications for this work have been in expert systems, where the promise of a model that can update itself as data becomes available, has generated intense interest from the artificial intelligence community (Charniak, 1991, Kornfeld, 1991). However, the application of this work to a broader range of discrete data problems has been largely overlooked.

The purpose of this article is to show how the Bayesian graphical framework unifies and greatly simplifies many standard discrete data problems such as Bayesian log linear modeling with either complete or incomplete data, model selection and accounting for model uncertainty, closed population estimation, multinomial estimation with misclassification, double sampling and database error prediction. This list by no means exhausts the possible applications.

This is a methodological article. Our objective is to demonstrate the diverse range of potential applications, alert the reader to an exciting new methodology and hopefully stimulate further development.

1.1 An Outline of the Basic Framework

At the risk of over-simplification we sketch the basic framework for the Bayesian analysis of graphical models with a simple medical example involving dichotomous variables:

> In recent years, Extracorporeal Shockwave Lithotripsy (ESWL) has become the treatment modality of choice for kidney stones (Kiely *et al*,1990). In the standard setup, the lithotripter operator first locates the stone via a real time ultrasound image. In the style of a video game, the operator then uses a joystick to identify the stone on the image and hundreds of high frequency shockwaves are focussed at the targeted location. Each individual shockwave passes harmlessly into the body at a separate location, but at the point of focus (hopefully the stone) sufficient energy is generated to disintegrate the stone. The quality of the ultrasound image affects the chance of disintegration. Subsequent clearance of the stone from the urinary tract (the desired outcome) is usually preceded by disintegration.

A possibly reasonable model for this situation is given in Figure 1. This directed graph represents the assumption that Clearance (C) and Ultrasound Image Quality (I) are conditionally independent given Disintegration (D). The joint distribution of the three variables factors accordingly:

$$\mathrm{pr}(I, D, C) = \mathrm{pr}(I)\mathrm{pr}(D \mid I)\mathrm{pr}(C \mid D). \tag{1}$$

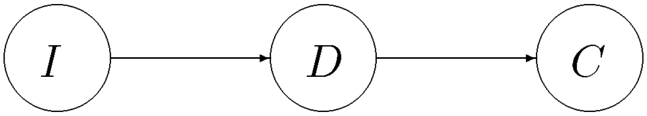


Figure 1: Lithotripsy: A Simple Discrete Graphical Model

At the time of treatment the primary quantity of interest is $\text{pr}(C \mid I)$, the probability of successful outcome given the quality of the ultrasound image (good/bad).

The approach to this problem pioneered by Spiegelhalter (1986) is based entirely on subjective expert knowledge. To fully specify the joint distribution in (1), five probabilities must be elicited:

$$\text{pr}(C \mid D), \text{pr}(C \mid \overline{D}), \text{pr}(D \mid I), \text{pr}(D \mid \overline{I}) \quad \text{and} \quad \text{pr}(I) \tag{2}$$

where the vinculum denotes negation. The calculation of $\text{pr}(C \mid I)$ could now proceed by writing down the eight probabilities of the joint distribution and marginalising over D. In general however, this may not be possible because the storage requirements for the joint distribution become prohibitively expensive as the number of variables increases. Spiegelhalter (1986) described a method for converting the directed representation in Figure 1 to an undirected representation corresponding to a specific log linear model. The calculation of arbitrary conditional probabilities can then proceed via a series of local calculations thereby sidestepping the need to store the full joint distribution. See Dawid (1992) and Lauritzen (1992) for recent discussions of similar algorithms.

An obvious development of the above framework is to update knowledge about the model parameters as data accumulate thereby providing an extension from probabilistic reasoning to statistical analysis. The use of point estimates for the probabilities in (2) precludes the possibility of such updating so instead we elicit prior distributions for these quantities. In effect the probabilities become random variables and can be added to the graph as in Figure 2. Within this framework Spiegelhalter and Lauritzen (1990) show how independent beta distributions placed on these probabilities can be updated locally to form the posterior as data becomes available. This provides an attractive strategy for Bayesian analysis of discrete data. The graph provides a powerful medium with which to communicate model assumptions and derive model properties. Informative prior distributions can realistically be elicited in terms of quantities that are well understood rather than, for example, the cryptic "u"-parameters of log linear models (Bishop *et al.*, 1975). Furthermore, the required computations are straightforward.

In later sections we extend this framework and apply it to a variety of problems. Some common themes will be apparent across these applications:

First, the importance of recognising and incorporating model uncertainty has been acknowledged by many authors. Hodges (1987), Raftery (1988b), and Draper (1993) argue

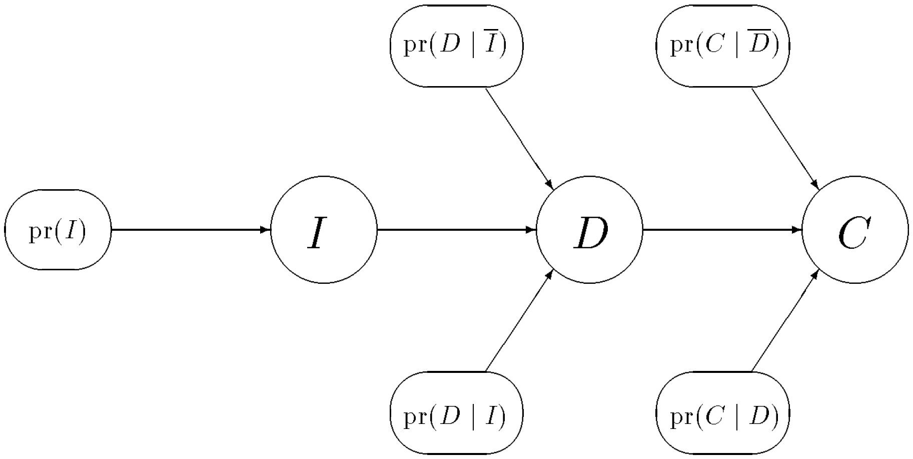


Figure 2: Lithotripsy: Bayesian Graphical Model

convincingly that ignoring model uncertainty can lead to underestimation of the uncertainty about quantities of interest. In the context of the lithotripsy example, inference about the relationship between image quality and clearance could be greatly affected by the addition of a link from I to C. Thus uncertainty about the conditional independence of C and I given D should be accounted for in subsequent inference. A complete Bayesian solution to this problem involves averaging over all possible models when making inferences about quantities of interest, much as one would integrate out a nuisance parameter in a hierarchical model. Indeed Hodges (1987) comments that "what is clear is that when the time comes for betting on what the future holds, one's uncertainty about that future should be fully represented, and model [averaging] is the only tool around." In many applications, however, because of the size of the model space and awkward integrals, this averaging will not be a practical proposition, and approximations are required. Draper (1993) describes "model expansion": averaging over all plausible models in the neighborhood of a "good" model. Madigan and Raftery (1991) describe an approach for Bayesian graphical models that involves seeking out the most plausible models and averaging over them. Raftery (1992) applies this to structural equation models. Here, we propose a Markov chain Monte Carlo approach which provides a workable approximation to the complete solution. The point is that with Bayesian graphical models, accounting for model uncertainty is entirely possible. This will be demonstrated in later applications.

Second, in several of the applications we consider, the presence of missing data and/or latent variables produces ostensibly insurmountable analytic obstacles. Such complexity frequently rules out the consideration of larger models involving many covariates and other generalisations. We will show how Bayesian graphical models coupled with Markov chain Monte Carlo techniques provide a conceptually simple approach to such problems and greatly extend the range of possible applications.

Finally, Bayesian graphical models provide an exciting opportunity to implement the complete Bayesian paradigm. The elicitability of *informative* prior distributions motivates many of the constructions we present in later sections.

In summary, there are many advantages to analysing discrete data with Bayesian graphical models:

- Most model assumptions are entirely transparent when a graphical representation of the model is used (Lange, 1992);
- Bayesian graphical models and attendant modeling strategies provide a unified and conceptually simple framework for a diverse range of applications;
- Model uncertainty can be accounted for in a straightforward fashion;
- Missing data and latent variables are catered for;
- Informative subjective knowledge can realistically be elicited and incorporated.

The primary disadvantage is increased computational complexity. However, with the advent of Markov chain Monte Carlo methods for Bayesian analysis and the widespread availability of immense computing power, this problem is somewhat mitigated.

1.2 Plan

In the next section, we define graphical models and describe more fully the Bayesian framework sketched above. The closed population estimation problem is presented in Section 3. In Section 4 we consider a simple application concerned with the estimation of multinomial probabilities subject to misclassification. Next we consider a range of double sampling problems and in Section 6 we re-examine some recent work concerning the estimation of errors in databases. Finally we discuss possible extensions of this work and other potential applications.

2 An Outline of the Technical Framework

2.1 Independence Graphs and Factorisations

Graphical models are a class of statistical models defined by collections of conditional independencies which can by represented by a graph (see Appendix I for a summary of the graph terminology we use). We will only consider graphs that are either directed and acyclic or undirected in what follows, although combinations of the two have also been studied—see for example, Whittaker (1990). In either case, each node in the graph will correspond to a random variable $X_v, v \in V$ taking values in a sample space $\mathcal{X}_v$.

In the directed case (see for example Figure 1), the parents pa(v) of a node v are those nodes from which edges point into v. These parents are taken to be the only direct influences on v, and thus, v is independent of its non-descendents given its parents.

This property implies a factorisation of the joint distribution of $X_v, v \in V$, which we denote by $\mathrm{pr}(V)$, given by:

$$\mathrm{pr}(V) = \prod_{v \in V} \mathrm{pr}(v \mid \mathrm{pa}(v)). \tag{3}$$

The class of models which can be defined in this way were introduced by Kiiveri *et al.* (1984) and are a subclass of their recursive causal models. Determining conditional independencies in large directed graphs can be difficult. However, Lauritzen *et al.* (1990) show that for sets A, B and $S \subset V$, A and B are conditionally independent given S, whenever A and B are separated by S in a "moralized" undirected graph containing $A \cup B \cup S$ and their ancestors. A moralized graph is formed by placing edges between nodes which share a child and then dropping the edge directions.

In the undirected case, we take each node to be conditionally independent of all others given its neighbours. For a more detailed exposition of Markov properties with respect to directed and undirected graphs, we refer the reader to Lauritzen *et al.* (1990).

In the case where the random variables $(X_v), v \in V$ are all discrete, the class of models defined by the *undirected* graphs are a subclass of the hierarchical log linear models where the cliques of the graph correspond to the maximal terms in the log linear model.

In what follows we will make extensive use of "decomposable models" for which the underlying undirected graph is chordal. These are the "closed-form" log linear models for which parameters can be estimated without recourse to iterative methods. The key property of such models is a simple factorisation of the joint density:

$$\mathrm{pr}(V) = \frac{\prod_{i=1}^{n} \mathrm{pr}(C_i)}{\prod_{i=2}^{n} \mathrm{pr}(S_i)} \tag{4}$$

where $C_1, \ldots, C_n$ is a so-called "perfect" clique ordering and $S_2, \ldots, S_n$ are the corresponding clique separators. The simplicity of such decomposable models has been exploited in a number of contexts—see for example Lauritzen and Spiegelhalter (1988), Dawid and Lauritzen (1993), Madigan and Mosurski (1990, 1991) and Madigan and Raftery (1991).

2.2 Bayesian Framework for Directed Graphical Models

Here we describe the Bayesian framework for directed graphical models. Consider a directed graphical model for a set of discrete random variables $X_v, v \in V$. The assumptions of the model imply that the joint distribution of $X_v, v \in V$ is given by Equation (3).

Spiegelhalter and Lauritzen (1990) introduced a parametrisation for $\mathrm{pr}(v|\mathrm{pa}(v))$ whereby the relationship between a node v and its parents $\mathrm{pa}(v)$ is fully specified by a possibly vector-valued parameter $\theta_v \in \Theta_v$. This leads to a conditional distribution for V:

$$\mathrm{pr}(V|\theta) = \prod_{v \in V} \mathrm{pr}(v|\mathrm{pa}(v), \theta_v).,\tag{5}$$

where θ has components θ_v corresponding to each node $v \in V$. For the lithotripsy example of Section 1, we have $\theta_C = \{\mathrm{pr}(C \mid D), \mathrm{pr}(C \mid \overline{D})\}, \theta_D = \{\mathrm{pr}(D \mid I), \mathrm{pr}(D \mid \overline{I})\}$ and $\theta_I = \{\mathrm{pr}(I)\}$.

Spiegelhalter and Lauritzen (1990) make two key assumptions which greatly simplify subsequent analysis. The first assumption is that of *global independence* whereby the parameters θ_v are assumed mutually independent *a priori.* This assumption alone allows us to calculate the likelihood for a single case:

$$\mathrm{pr}(v) = \int \mathrm{pr}(v, \theta) d\theta = \int \prod_v \mathrm{pr}(v|\mathrm{pa}(v), \theta_v)\mathrm{pr}(\theta_v)d\theta_v = \prod_v \mathrm{pr}(v|\mathrm{pa}(v))$$

where

$$\mathrm{pr}(v|\mathrm{pa}(v)) = \int \mathrm{pr}(v|\mathrm{pa}(v), \theta_v)\mathrm{pr}(\theta_v)d\theta_v.$$

The second assumption is that of *local independence* whereby components of θ_v corresponding to the elements of the state space of $\mathrm{pa}(v)$ are assumed to be mutually independent *a priori.* Both of these assumptions were embodied in the lithotripsy example of Figure 2, where, for instance, we have that $\mathrm{pr}(I)$ is independent of $\mathrm{pr}(D \mid I)$ (global independence) and $\mathrm{pr}(D \mid I)$ is independent of $\mathrm{pr}(D \mid \overline{I})$ (local independence).

Now consider a conditional probability distribution $\mathrm{pr}(v|\mathrm{pa}(v)^+, \theta_v^+) = \theta_v^+$ for a specific state $\mathrm{pa}(v)^+$ of $\mathrm{pa}(v)$. We assume that θ_v^+ has a Dirichlet distribution $\mathcal{D}[\lambda_1^+, ..., \lambda_k^+]$ where k is the number of states of v (alternative parametrisations are also considered by Spiegelhalter and Lauritzen, 1990). This prior is conjugate with multinomial sampling, and it follows that:

$$\mathrm{pr}(v|\mathrm{pa}(v)^+) = \lambda_v^+ / \sum_i \lambda_i^+$$

thereby providing a simple method for calculating the likelihood.

If we observe one data case where v is in state j and the parent state is $\mathrm{pa}(v)^+$, the posterior distribution of θ_v^+ is given by:

$$\theta_v^+ | v \sim \mathcal{D}[\lambda_1^+, ..., \lambda_j^+ + 1, ..., \lambda_k^+].$$

In general, the posterior distributions are found by incrementing each parameter λ_j^+ by the number of cases with that configuration of v and $\mathrm{pa}(v)$. If the data are complete, updating each component of θ in this fashion preserves local and global independence.

2.3 Bayesian Framework for Undirected Decomposable Graphical Models

Following Dawid and Lauritzen (1993), we consider a decomposable model M for a set of random variables $X_v, v \in V$. Let $\mathcal{I} = \prod_{v \in V} \mathcal{X}_v$ denote the set of possible configurations of X. Denote by $\theta(i)$ the probability of a state $i \in \mathcal{I}$. Then $\theta(i)$ is determined by the clique marginal probability tables $\theta_C, C \in \mathcal{C}$ where $\mathcal{C}$ denotes the set of cliques of M:

$$\theta(i) = \frac{\prod_{C \in \mathcal{C}} \theta_C(i_C)}{\prod_{S \in \mathcal{S}} \theta_S(i_S)}, i \in \mathcal{I}.$$

$\mathcal{S}$ denotes the system of clique separators in an arbitrary perfect ordering of $\mathcal{C}$.

For each clique $C \in \mathcal{C}$, let $\mathcal{D}(\lambda_C)$ denote the Dirichlet distribution for θ_C with density

$$\pi(\theta_C|\lambda_C) \propto \prod_{i_C \in \mathcal{I}_C} \theta_C(i_C)^{\lambda_C(i_C)-1},$$

where $\lambda_C(i_C) > 0$ for all $i_C \in \mathcal{I}_C$.

Now let us suppose that the collection of specifications $\mathcal{D}(\lambda_C), C \in \mathcal{C}$ are constructed in such a way that for any two cliques C and D in $\mathcal{C}$ we have:

$$\lambda_C(i_{C\cap D}) = \lambda_D(i_{C\cap D}); \tag{6}$$

that is, if the cliques C and D overlap, then the parameters λ_C and λ_D are such that each implies the same marginal distribution for $\theta_{C\cap D}$. Dawid and Lauritzen (1993) have shown that there exists a unique "hyper-Dirichlet" distribution for θ over M such that θ_C has the marginal density $\mathcal{D}(\lambda_C)$ for all $C \in \mathcal{C}$.

In practice, one would construct a hyper–Dirichlet distribution by first identifying a perfect ordering of the cliques $\{C_1, \ldots, C_n\}$. Place a Dirichlet distribution $\mathcal{D}(\lambda_{C_1})$ on θ_{C_1}; next place a Dirichlet distribution $\mathcal{D}(\lambda_{C_2})$ on θ_{C_2}, with parameters constrained by (6) and realizations constrained so that $\theta_{C_1 \cap C_2}$ is identical for θ_{C_1} and θ_{C_2}. For each subsequent clique C_i, place a Dirichlet on θ_{C_i} such that the parameters and the realizations of that distribution are consistent with those specified for the previous cliques.

This prior distribution is conjugate with multinomial sampling. Simple expressions for posterior distributions and likelihoods are provided in Dawid and Lauritzen (1993).

2.4 Directed vs Undirected Graphical Models

In general, probability distributions can have conditional independence properties more complex than can be represented with either an undirected or directed graph; see Pearl (1988). However, it is always possible to provide a graph for a probability distribution such that any independence assumptions present in the graph are true for the distribution–Pearl (1988) calls such a graph an I-map. A trivial example of this would be a fully connected undirected graph, which makes no independence assumptions at all. Thus, we can always find a graphical model which makes no false independence assertions, although it may have more parameters than would be strictly necessary. If the graph is such that it is an I-map for a distribution, and every independence relationship in the distribution is represented in the graph, Pearl (1988) calls it a *perfect map* of the distribution.

Additionally, there are distributions such that there is an undirected graph that is a perfect map, but no directed graph that is a perfect map, and *vice versa.* These two types of graphs can express different kinds of relationships, which raises the question of which type should be used for any given problem.

In problems where some variables obviously are determined before others, or cause others, the directed graphs allow a direct representation of these assumptions. For example, if there is a relationship between the kidney stone disintegration (D) kidney stone clearance (C), it is certainly D which influences or causes or precedes C, and not the other way around; thus, an edge between them should point from D to C.

Undirected models, in contrast, are best suited to problems where the variables are determined simultaneously, or perhaps are both influenced by some variable which is not explicitly modeled. For example, it does not make sense to say that an individual's eye color influences or causes his or her hair color, or vice versa, and so a relationship between these variables is better represented as an undirected edge.

Many problems will include both kinds of relationships, motivating the use of graphs with both directed and undirected edges (Frydenberg, 1990). Currently we are extending the class of Bayesian graphical models to include such graphs.

2.5 Accounting for Model Uncertainty

A typical approach to data analysis is to initially carry out a model selection exercise leading to a single "best" model and to then make inference as if the selected model were the true model. However, as a number of authors have pointed out, this paradigm ignores a major component of uncertainty, namely uncertainty about the model itself (Breslow, 1991, Draper *et al.* (1987), Draper, 1993, Hodges, 1987, Moulton, 1991, Raftery, 1988b). As a consequence uncertainty about quantities of interest can be underestimated. For striking examples of this see Regal and Hook (1991), Draper (1993), Miller (1984), and York and Madigan (1992).

There is a standard Bayesian way around this problem. If Δ is the quantity of interest, such as a structural characteristic of the system being studied, a future observation, or the utility of a course of action, then its posterior distribution given data D is:

$$\mathrm{pr}(\Delta \mid D) = \sum_{k=1}^{K} \mathrm{pr}(\Delta \mid M_k, D)\mathrm{pr}(M_k \mid D). \tag{7}$$

This is an average of the posterior distributions under each of the models, weighted by their posterior model probabilities. In equation (7), $M_1, \ldots, M_K$ are the models considered, the marginal likelihood for model M_k is given by:

$$\mathrm{pr}(M_k \mid D) = \frac{\mathrm{pr}(D \mid M_k)\mathrm{pr}(M_k)}{\sum_{l=1}^{K} \mathrm{pr}(D \mid M_l)\mathrm{pr}(M_l)}, \tag{8}$$

where

$$\mathrm{pr}(D \mid M_k) = \int \mathrm{pr}(D \mid \theta, M_k)\mathrm{pr}(\theta \mid M_k)d\theta, \tag{9}$$

θ is the vector of cell probabilities, $\mathrm{pr}(\theta \mid M_k)$ is the prior for θ under model M_k, $\mathrm{pr}(D \mid \theta, M_k)$ is the likelihood, and $\mathrm{pr}(M_k)$ is the prior probability that M_k is the true model.

Furthermore, averaging over *all* the models in this fashion provides better predictive ability, as measured by a logarithmic scoring rule, than using any single model M_j (Madigan and Raftery, 1991).

However, as Breslow (1991) points out, implementation of the above strategy is difficult. There are two primary reasons for this: first, the integrals in (9) can in general be hard to compute, and second, the number of terms in (7) can be enormous.

We consider two approaches to this problem. Madigan and Raftery (1991) do not attempt to approximate (7) but instead, appealing to standard norms of scientific investigation, adopt a model selection procedure. This involves averaging over a much smaller set of models than in (7) and delivers a parsimonious set of models to the data analyst, thereby facilitating effective communication of model uncertainty. A second approach we propose here does involve approximating (7) with a Markov chain Monte Carlo method.

Before sketching the two approaches, we note that both involve repeated calculation of terms like:

$$\frac{\mathrm{pr}(M_0 \mid D)}{\mathrm{pr}(M_1 \mid D)} \tag{10}$$

where M_0 and M_1 are graphical models which differ by one link. In both the directed and undirected (decomposable) case these ratios can be calculated in a highly efficient manner entirely through local computations. For details, see Madigan and Raftery (1991).

Two basic principles underly the approach of Madigan and Raftery (1991). Firstly, they argue that if a model predicts the data far less well than the model which provides the best predictions, then it has effectively been discredited and should no longer be considered. Thus models not belonging to:

$$\mathcal{A}' = \left\{ M_k : \frac{\max_l\{\mathrm{pr}(M_l \mid D)\}}{\mathrm{pr}(M_k \mid D)} \leq C \right\}, \tag{11}$$

should be excluded from equation (7) where C is chosen by that data analyst. Secondly, appealing to Occam's razor, they exclude complex models which receive less support from the data than their simpler counterparts. More formally they also exclude from (7) models belonging to:

$$\mathcal{B} = \left\{ M_k : \exists M_l \in \mathcal{A}, M_l \subset M_k, \frac{\mathrm{pr}(M_l \mid D)}{\mathrm{pr}(M_k \mid D)} > 1 \right\} \tag{12}$$

and equation (7) is replaced by

$$\mathrm{pr}(\Delta \mid D) = \sum_{M_k \in \mathcal{A}} \mathrm{pr}(\Delta \mid M_k, D)\mathrm{pr}(M_k \mid D) \tag{13}$$

where

$$\mathcal{A} = \mathcal{A}' \backslash \mathcal{B}. \tag{14}$$

This greatly reduces the number of models in the sum in equation (7) and now all that is required is a search strategy to identify the models in $\mathcal{A}$. Two further principles underly the search strategy. Firstly, if a model is rejected then all its submodels are rejected. This is justified by appealing to the independence properties of the models. The second principle — "Occam's Window" — concerns the interpretation of the ratio of posterior model probabilities $\mathrm{pr}(M_0 \mid D)/\mathrm{pr}(M_1 \mid D)$. Here M_0 is one link "smaller" than M_1. The essential idea is shown in Figure 3: If there is evidence for M_0 then M_1 is rejected but to reject M_0 we require strong evidence *for* the larger model, M_1. If the evidence is inconclusive (falling in Occam's Window) neither model is rejected. Madigan and Raftery (1991) adopted $\frac{1}{20}$ for O_L and 1 for O_R.

These principles fully define the strategy. Typically the number of terms in (7) is reduced to fewer than 20 models and often to as few as two. Madigan and Raftery (1991) provide a detailed description of the algorithm and show how averaging over the selected models

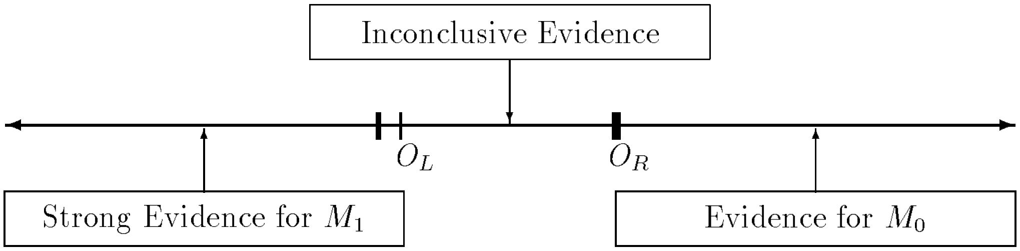


Figure 3: Occam's Window: Interpreting the log posterior odds

provides better predictive performance than basing inference on a single model in each of the examples they consider.

Our second approach is to approximate (7) using Markov chain Monte Carlo methods, such as in Metropolis *et al.* (1953) and Hastings (1970), generating a process which moves through model space. Specifically, let $\mathcal{M}$ denote the space of models under consideration. We can construct a Markov chain $\{M(t)\}, t = 1, 2, \ldots$ with state space $\mathcal{M}$ and equilibrium distribution $\text{pr}(M_i \mid D)$. Then for a function $g(M_i)$ defined on $\mathcal{M}$, if we simulate this Markov chain for $t = 1, \ldots, N$, the average:

$$\hat{G} = \frac{1}{N} \sum_{t=1}^{N} g(M(t)) \tag{15}$$

is an estimate of $E(g(M))$. Applying the ergodic theorem (see Breiman, 1968, or Chung, 1967) for finite irreducible Markov chains,

$$\hat{G} \rightarrow \mathbf{E}(g(M)) \; a.s. \text{ as } N \rightarrow \infty.$$

To compute (7) in this fashion set $g(M) = \text{pr}(\Delta \mid M, D)$.

To construct the Markov chain we define a neighbourhood $\text{nbd}(M)$ for each $M \in \mathcal{M}$ which is the set of models with either one link more or one link fewer than M and the model M itself. Define a transition matrix q by setting $q(M \rightarrow M') = 0$ for all $M' \notin \text{nbd}(M)$ and $q(M \rightarrow M')$ non–zero for all $M' \in \text{nbd}(M)$. If the chain is currently in state M, we proceed by drawing M' from $q(M \rightarrow M')$; if M' is "legal" (it contains no directed cycles in the directed case and is chordal in the undirected case) it is accepted with some positive probability chosen so that the process has the correct stationary distribution. Some possibilities for these acceptance probabilities are given by Hastings (1970).

The irreducibility of the transition matrix q is obvious in the directed case. For the decomposable case it follows from Lemma 5 of Frydenberg and Lauritzen (1989).

The choice of which approach to use — model selection or Markov chain Monte Carlo model composition — will depend on the particular application. The model selection procedure will be most useful when one is interested in making inferences about the relationships between the variables. Averaging over all models (by brute force or Monte Carlo) will be appropriate for making predictions or decisions when the posterior distribution of some quantity is of more interest than the nature of the "true" model. However, each approach is flexible enough to be used successfully for inference *and* prediction.

Madigan *et al.* (1993a) contrast the two approaches. In each of the three applications they consider the Monte Carlo approach provides better predictive ability than the Occam's window approach. However, either method provides improved predictive performance over inference based on any single model that might reasonably have been selected.

We note that similar approaches are suggested by Cooper and Herskovits (1992).

3 Bayesian Graphical Models for Closed Population Estimation

We now introduce the first of several applications demonstrating the utility of Bayesian graphical models.

3.1 Introduction

One approach to estimating the size of a closed population is to use several methods to "capture" individuals in the population. Although these might be actual physical captures, here we will consider a capture to be the occurrence of a person's name on an administrative list. If it is possible to uniquely identify the individuals or their capture histories, then the data can be represented as a contingency table with one dimension for each capture method. The count for each cell gives the number of individuals with a particular capture history. Of course, the count for the cell in which individuals were not caught by any of the methods is unknown. The goal of the analysis is to estimate the number of individuals in this cell and thence in the population.

For example, consider estimating the rate at which the birth defect *spina bifida* occurs. Hook *et al* (1980) gathered records on persons born in upstate New York between 1969 and 1974 with this defect from birth certificates (B), death certificates (D), and medical rehabilitation records (R). These different records were compared, and each individual with the defect was classified as to whether or not they were found in each list. The data is given in Table 1, where a value of 0 indicates that an individual was not found in that list, and a 1 indicates that he or she was found. A total of 626 individuals were found in the three record systems considered, out of a total of 863,143 live births; the question is, how many more individuals were missed by all three?

This is sometimes called the multiple record systems (MRS) problem and is related to experiments where animals are physically captured or tagged (El-Khorazaty *et al*, 1977). The essential difference between the two problems is that there will almost always be dependence between some of the lists in an MRS because of relationships between the administrative systems and heterogeneity in the population. In contrast, when dependence between captures is modeled in the capture-recapture literature it is usually done in a simple sequential manner, with one capture probability for individuals that have been captured before and another for those who have not yet been captured (Wolter, 1986; Pollock and Otto, 1983; Otis *et al*

	$R = 0$	$R = 1$
$B = 0, D = 0$	?	60
$B = 0, D = 1$	49	4
$B = 1, D = 0$	247	112
$B = 1, D = 1$	142	12

Table 1: Spina Bifida data

1978). For an MRS, this approach will not be as useful, because the different administrative systems may be operating simultaneously and the dependencies between them may follow some more general pattern.

Log linear models provide a flexible approach to this problem, in which dependence between lists is explicitly modeled (Fienberg, 1972, Bishop *et al*, 1975, Hook *et al*, 1980). However, as highlighted in Section 2.5, to ignore model uncertainty in this context is to ignore an important aspect of uncertainty in predicting the population size. Inadequacies of model selection routines for MRS problems have been illustrated by Regal and Hook (1991), and are discussed in more detail by York and Madigan (1992). Similar complaints about model selection with standard capture–recapture models can be found in Merkins and Anderson (1988).

Undirected decomposable Bayesian graphical models, as described in Section 2 provide a flexible model class for this problem, facilitating the incorporation of prior expert knowledge and accounting for model uncertainty.

Denoting the total population size by N, and following the notation of Section 2, our objective is to evaluate the posterior distribution of N, given the observed data, D:

$$\mathrm{pr}(N \mid D) = \sum_{k=1}^{K} \mathrm{pr}(N \mid M_k, D)\mathrm{pr}(M_k \mid D). \tag{16}$$

We assume *a priori* that N is independent of the model M_k, and thus (16) can be written as:

$$\mathrm{pr}(N \mid D) = \sum_{k=1}^{K} \mathrm{pr}(D \mid M_k, N)\mathrm{pr}(M_k)\mathrm{pr}(N)/\mathrm{pr}(D), \tag{17}$$

where:

$$\mathrm{pr}(D \mid M_k, N) = \int \mathrm{pr}(D \mid \theta, M_k, N)\mathrm{pr}(\theta \mid M_k)d\theta. \tag{18}$$

Here, θ is the vector parameter of probabilities which define M_k and is assumed to be independent of N. York and Madigan (1992) provide formulae for $\mathrm{pr}(D \mid \theta, M_k, N)$ and $\mathrm{pr}(D)$, and explore the consequences of various prior assumptions for $\mathrm{pr}(M_k)$ and $\mathrm{pr}(N)$ and for $\mathrm{pr}(\theta \mid M_k)$, in the context of several examples.

The component distributions of $\mathrm{pr}(\theta \mid M_k)$ all involve the probability of capture on one or more lists. One practical difficulty that arises is that the structure of this prior distribution depends on M_k; recall that a Dirichlet distribution is required for each clique in the graph of M_k. This necessitates the elicitation of a different prior distribution for every model.

York and Madigan (1992) describe a pragmatic solution to this problem. Essentially, our approach is to elicit a prior distribution for $\text{pr}(\theta \mid M_{k'})$, where $M_{k'}$ is a model with high prior probability, chosen for convenient elicitation. Prior distributions for θ under all the other models are then derived from $\text{pr}(\theta \mid M_{k'})$, via a simple information theoretic argument. See Spiegelhalter *et al.* (1993) for a similar approach.

Elicitation for undirected graphs can be difficult. In the context of the spina-bifida example above, a link from say, B to D, could require the elicitation of a prior distribution over the 2×2 table spanned by B and D. Madigan and Raftery (1991) describe in detail an alternative approach whereby the prior distribution is elicited in the context of a directed graph. Subsequently, prior distributions for the components of θ implied by the undirected model are derived.

The key point is that the elicitation of informative prior distributions, while not without its difficulties, *is* possible. This is not the case for the equivalent distributions in the conventional log-linear framework.

3.2 Example : Spina Bifida

The results of a Bayesian graphical model analysis of the spina bifida example of Table 1 are given in Table 2 and Figure 4. In the figure, the value of

$$P(N \mid D, M_k)P(M_k \mid D)$$

is plotted for any model M_k with non-negligible posterior probability. These curves show both the shape of the posterior distribution of N for particular models, and, by the area beneath them, their relative contribution to the overall posterior distribution. The sum of these curves gives the full posterior, averaged over all models. In this analysis, uniform prior distributions were adopted for the components of θ under the largest model, all models were assumed equally likely *a priori* and an informative prior distribution, based on historical data, was adopted for N. For details we refer the reader to York and Madigan (1992).

We note that the posterior distributions for N under the three leading models are centered at different locations, and estimation of N conditional upon one model would depend a great deal upon the particular model chosen. Averaging over models, on the other hand, gives us a single posterior distribution that accurately reflects our uncertainty about the correct model. A detailed coverage analysis is described in York and Madigan (1992) which shows that model averaging provides prediction intervals which are much better calibrated than those based on a single model.

The posterior means and standard deviations for the probability that an individual will be found via any particular list are given in Table 3. It is awkward to come up with such "efficiency" estimates in the conventional log-linear modeling framework. In contrast, the methods described here, directly and easily produce efficiency estimates for each list.

If we use our estimate of N to compute a prevalence rate for spina bifida for the population of all live births, we arrive at an estimate of 0.847 per 1000 births, with 2.5th and 97.5th posterior percentiles being 0.790, 0.923. In comparison to the estimate of 0.725 per thousand if we assume that no cases were missed, there is substantial evidence that more than one case per ten thousand is missed; and this is for a population count based on three separate lists.

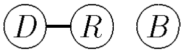
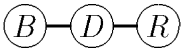
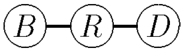
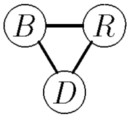
	Posterior		
Model	Prob.	$\hat{N}$	2.5%, 97.5%
	0.373	731	(701, 767)
	0.301	756	(714,811)
	0.281	712	(681,751)
	0.036	697	(628,934)
Model Averaging	—	731	(682,797)

Table 2: Summaries of the posterior distributions of N for the spina bifida data for all models with posterior probability greater than 0.01. $\hat{N}$ is a Bayes estimate, minimizing a relative squared error loss function

	Posterior	Posterior
List	Mean	Std. Dev
Birth Certificate	0.699	0.032
Death Certificate	0.284	0.020
Medical Records	0.258	0.019

Table 3: Posterior mean and standard deviation for the probability that a given list will correctly identify an individual with spina bifida

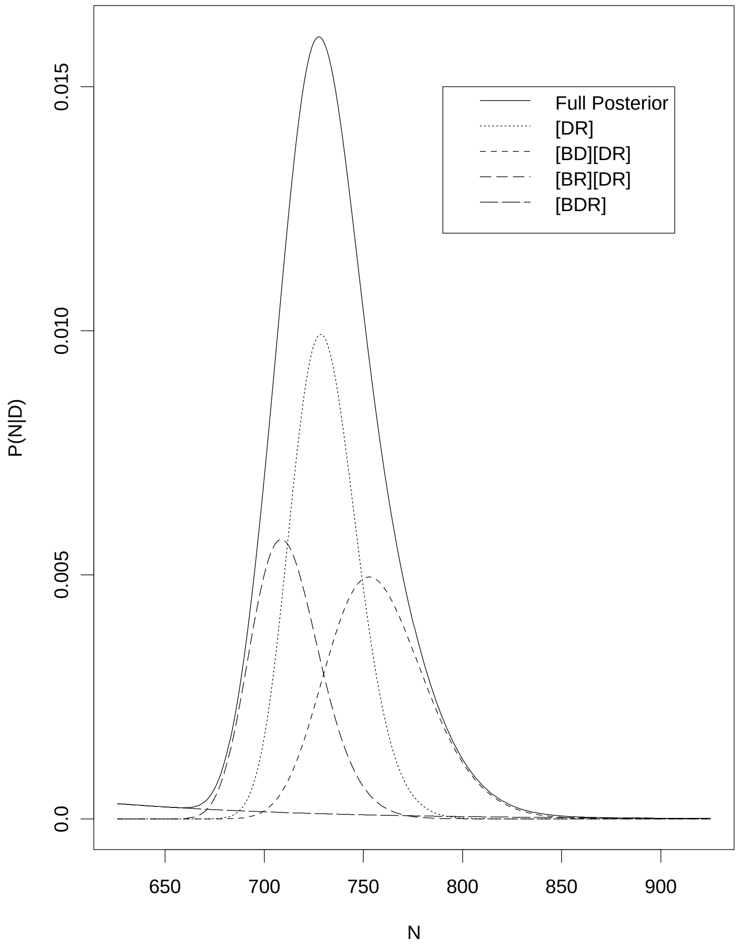


Figure 4: Posterior distribution for the number of cases of spina bifida for different models. “Full Posterior” shows the posterior distribution averaged over all the decomposable Bayesian graphical models

The estimate of 0.699 for the efficiency of birth certificates alone indicates that around 30% of the total cases would be overlooked if that registry were the sole source of information.

3.3 Example : Spina Bifida with Covariate

One of the benefits of using Bayesian graphical models for discrete data analysis is the comparative ease with which models can be expanded. To illustrate this point we consider the addition of a covariate, race, in the spina bifida example. The data presented in Table 4 are from Hook *et al.* (1980). In addition to the data in the table, there are 5 individuals for whom we have no information on race. These five individuals had the following values for (B, D, R) : $(1, 0, 1)$, $3 \times (1, 0, 0)$, and $(0, 1, 0)$. We compute the posterior distribution for N by summing over all 2^5 possible values of race for these 5 incomplete cases, as well as summing over the possible races of the unobserved individuals.

	Whites		Blacks & Others	
	$R = 0$	$R = 1$	$R = 0$	$R = 1$
$B = 0, D = 0$	?	52	?	8
$B = 0, D = 1$	45	3	3	1
$B = 1, D = 0$	230	107	14	4
$B = 1, D = 1$	134	12	8	0

Table 4: Spina Bifida data, by Race

The posterior distribution for N is displayed in Figure 5. Features of the posterior distribution and the models which make the greatest contribution are given in Table 5. Most of the models which had high posterior probability in the previous analysis still have high probability here. The notable additions are several models with an interaction between birth certificates B and ethnicity, E. The posterior probability of a link between the two is 0.69; the posterior probability of any other link with E is less than 0.10. The models with a link to ethnicity tend to have higher estimates for the non-white population than the other models, indicating that the administrative lists seem to be missing proportionately more non-whites than whites. However, inclusion of this information does not cause a change in the overall estimate of the population size.

Efficiencies of the various lists, broken down by race, are given in Table 6: birth certificates are considerably less effective in identifying spina bifida cases in non-whites.

3.4 Why Bayesian graphical models?

Bayesian graphical models allow for flexible modeling of inter-list dependencies together with an effective medium to communicate these dependencies, i.e., a graph. Furthermore, informative expert knowledge can be expressed directly in terms of well-understood quantities, distributions for other quantities of interest such as list efficiencies are easily computed, inclusion of covariates is straightforward (including missing values), and crucially, model uncertainty can be effectively communicated and accounted for.

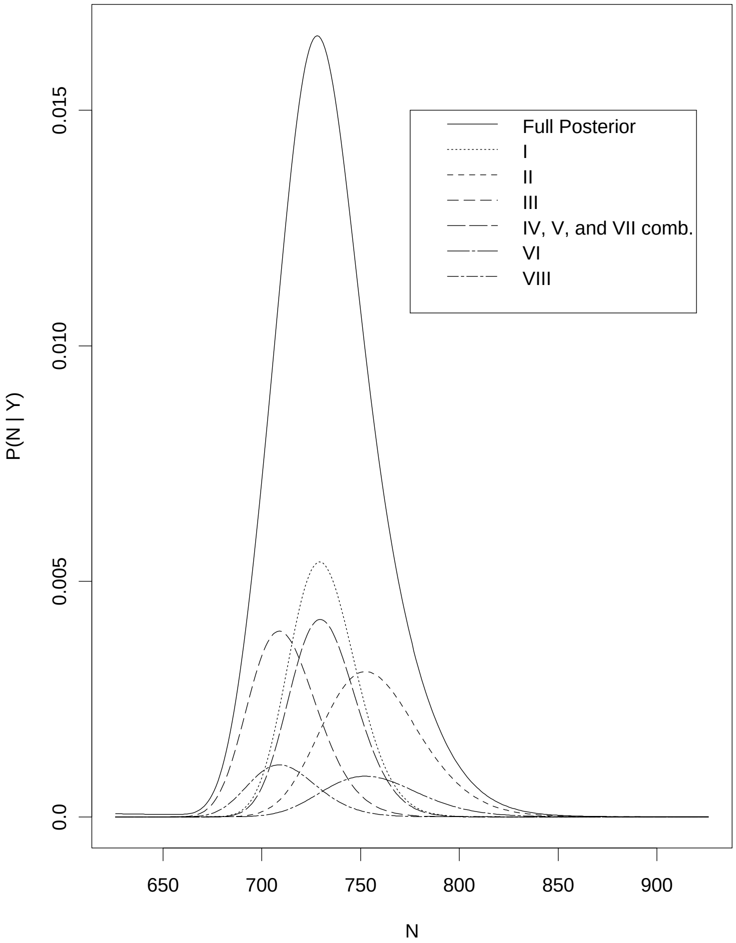


Figure 5: Posterior distribution for the number of cases of spina bifida, with race as a covariate. Labels refer to those in Table 5; the scaled posterior distributions for models IV, V, and VII have been added together since they share nearly the same posterior probability and shape.

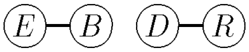
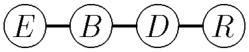
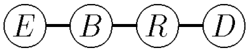
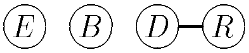
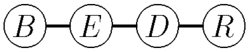
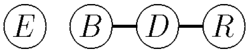
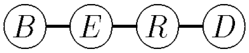
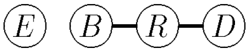
Label	Model	Posterior Prob	$\hat{N}$	(2.5, 97.5)	Whites $\hat{N}$	Non-whites $\hat{N}$
I		0.223	731	(701,767)	683	48
II		0.185	756	(714,811)	699	56
III		0.168	712	(681,751)	660	50
IV		0.062	731	(701,767)	683	48
V		0.061	732	(702,769)	677	54
VI		0.052	756	(714,811)	706	49
VII		0.050	731	(701,767)	678	52
VIII		0.047	712	(681,751)	665	46
	Model Avg	—	731	(689,794)	679	51

Table 5: Summaries of the posterior distributions of N for the spina bifida data for all models with posterior probability greater than 0.01 with race as a covariate.

	Whites		Non-whites	
List	Mean	Std. Dev	Mean	Std. Dev
Birth Certificate	0.710	0.033	0.565	0.107
Death Certificate	0.285	0.020	0.282	0.031
Medical Records	0.259	0.019	0.263	0.031

Table 6: Posterior mean and standard deviation for the probability that a given list will correctly identify an individual with spina bifida

The restriction to decomposable models may be a real concern for some applications. For the spina-bifida example, York and Madigan (1992) show that inclusion of the non-decomposable no-third-order-interaction model has little impact on the results.

This methodology could be applied to capture-recapture models as well. The sequential nature of those captures make directed graph representations more natural–see for example Rodrigues *et al.* (1988).

4 Multinomial Misclassification

In this section we present a simple application of Bayesian graphical models with latent variables from the field of systematic musicology. There are many approaches one could adopt for this problem. The advantages of the Bayesian graphical model approach are that firstly, the conceptual simplicity of the framework allows for the elicitation of informative priors and secondly, unlike more conventional approaches, Bayesian graphical models can easily be scaled up to include covariates and alternative sampling schemes.

Our application concerns music expectancy, a psychological construct which has been of interest to musicians since the early part of this century (Bissell, 1921). Music expectancy is defined as the cognitive awareness of a future event to come in music as we listen, an awareness not only of the nature of the event to come, but also of when the event will occur (Carlsen *et al.*, 1992). Narmour (1990) has postulated certain patterns that musical expectancy should exhibit; these theories have intensified interest in the subject. Carlsen (1981) and Unyk and Carlsen (1987) reported analyses of large music expectancy data sets. We set out to re-analyse this data in the light of Narmour's new work, and to assess what level of support the data provided for his theories (Madigan *et al.*, 1992).

The melodic expectancy studies of Carlsen all utilised the so-called "production response": participants in those studies were presented with the 25 two-note melodic beginnings possible within the octave (12 ascending, 12 descending and the unison). They were instructed to consider that interval as the beginning of an interrupted melody and were asked to sing immediately (in tempo) the expected continuation of the melody as if it had not been interrupted. The data is represented as a 25×25 table of counts representing the 25 melodic beginnings and the 25 melodic continuations within an octave (less than 1% of the melodic continuations were outside an octave). The particular data set we consider was collected in the U.S.A. and contains 12,262 data points.

Narmour's "Implication-Realization" model consists of a template within this table where expectancies are postulated to exist—see Figure 6.

Our initial effort to contrast the data with Narmour's model was to construct an equivalent empirical model by simply thresholding the data. Cells with fewer than a certain level of counts were deemed to be 'outside' and the remainder 'inside' an empirical template. However, the arbitrariness of the threshold was problematic. For a significant number of cells it was unclear whether they represented genuine expectancies and should therefore be included in the template, or whether the counts were spurious and recorded in error.

Further consideration suggested that such errors could arise in two ways: firstly, singer error (SE) whereby the subject could fail to produce the note they had intended to sing, and secondly, listener error (LE) whereby the listener transcribing the subject's sung response

Melodic Continuation

Melodic Beginning	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12
-12																									
-11						●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●				
-10							●	●	●	●	●	●	●	●	●	●	●	●	●	●					
-9								●	●	●	●	●	●	●	●	●	●	●	●						●
-8									●	●	●	●	●	●	●	●	●	●						●	●
-7										●	●	●	●	●	●	●	●						●	●	●
-6																									
-5	●	●	●	●	●	●	●	●	●	●	●					●	●	●	●	●					
-4	●	●	●	●	●	●	●	●	●	●	●	●			●	●	●	●	●						
-3	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●	●	●							
-2	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●								
-1	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●									
0											●	●	●	●	●										
1										●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
2									●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
3								●	●	●	●	●		●	●	●	●	●	●	●	●	●	●	●	●
4							●	●	●	●	●			●	●	●	●	●	●	●	●	●	●	●	●
5						●	●	●	●	●					●	●	●	●	●	●	●	●	●	●	●
6																									
7	●	●	●						●	●	●	●	●	●	●	●									
8	●	●						●	●	●	●	●	●	●	●	●	●								
9	●						●	●	●	●	●	●	●	●	●	●	●	●							
10						●	●	●	●	●	●	●	●	●	●	●	●	●	●						
11					●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●					
12																									

Figure 6: Narmour's Model. The "•"'s represent melodic continuations that are predicted to occur for each melodic beginning. No predictions are made for the octave and tritone melodic beginnings (-12, -6, 6, and 12).

could make an error. For a single melodic beginning this sugegsts the graphical model of Figure 7. Here, LE and SE are binary variables indicating whether or not a listener error or a singer error have occurred, $\mathrm{MC_T}$ is the true unobservable melodic continuation and $\mathrm{MC_F}$ is the recorded (fallible) melodic continuation.

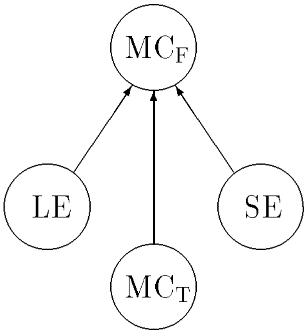


Figure 7: Initial Graphical Expectancy Model

This model requires the elicitation of prior distributions for $\mathrm{pr(MC_F \mid MC_T,LE,SE)}$, pr(LE), pr(SE) and $\mathrm{pr(MC_T)}$. Our approach was to elicit measures of location for these quantities and chose the scale to give diffuse Dirichlet priors. A Jeffreys prior was used for $\mathrm{pr(MC_T)}$.

The model of Figure 7 embodies the assumption that LE, SE and $\mathrm{MC_T}$ are mutually independent. However, it was felt that singer errors were more likely to occur for large intervals (tritone or larger) than smaller intervals. Concerning listener error, evidence presented in Unyk and Carlsen (1987) suggested that transcription errors were more likely to occur when the melodic beginning has high expectancy generating strength (e.g. $C - D$ which generates an expectancy of either E or C in the vast majority of subjects) and this expectancy is violated (e.g. $C - D$ followed by $G^\sharp$). The adjusted Bayesian graphical model reflecting these dependencies is given in Figure 8 and the required prior distributions, i.e. $\mathrm{pr(MC_F \mid MC_T,LE,SE)}$, $\mathrm{pr(LE \mid MC_T)}$, $\mathrm{pr(SE \mid MC_T)}$, and $\mathrm{pr(MC_T)}$, were easily elicited. For details, we refer the reader to Carlsen *et al.* (1992).

What we are interested in is of course the distribution of $\mathrm{MC_T}$ and in particular, the probability that all the counts in a given cell are spurious. A Gibbs sampling technique was employed to estimate these probabilities using the data augmentation idea of Tanner and Wong (1987)—see also Smith and Roberts (1993). A complication that arises here is that a Gibbs sampler Markov chain defined on the model of Figure 8 will not be irreducible—it is not possible to get from a state with neither listener error nor singer error to a state with errors updating just one variable at a time. A simple solution to this problem is to periodically update *two* variables, i.e. an error variable and an melodic continuation variable, simultaneously.

The resulting empirical model is shown in Figure 9. Cells marked O had no observations and are deemed to be 'outside' the template. Cells marked I, never had a zero count in 5,000 iterations of the Gibbs sampler and are deemed to be 'inside' the template. Cells marked

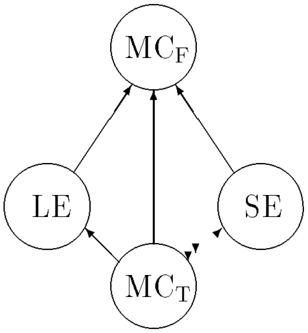


Figure 8: Graphical Expectancy Model

F for 'fuzzy' did have non-zero counts in the dataset but the probability that they are all spurious is non-zero.

Comparing Figure 9 with Narmour's I-R model (Figure 6), it is clear that there are substantial areas of disagreement between the two models. For a detailed comparison of the two models see Carlsen *et al.* (1992).

The method we have adopted here is a form of multinomial smoothing. A similar approach is suggested in Titterington (1985). Bayesian graphical models provide a practical method for doing this and allow for incorporation of expert knowledge expressed in terms of readily understood quantities. This is the chief advantage of the Bayesian graphical model-based approach in this application and provides a method to carry out "knowledge-based" smoothing as against the rather more arbitrary kernel-based methods.

5 Double Sampling

5.1 Introduction

Suppose you are presented with the following task: estimate the proportion of newborns born with jaundice nationwide. The data to hand consists of records of 500 births where it is recorded at birth by the midwife or gynecologist whether or not the child is jaundiced. This classification however is only based on a visual inspection of the child and may be incorrect. For a random subsample of 100 of the births highly accurate (but expensive) pathology tests are also available. Fictituous data are presented in Table 7 where D_F indicates the child is jaundiced according to the "fallible" visual inspection and D_T indicates jaundice according to the true or "infallible" pathology test (assumed here to be without error).

There are two obvious ways to estimate the required proportion. Just using the infallible pathology data gives an estimate of 0.68 with a standard deviation of 0.047. Alternatively, using just the visual data gives an estimate of 0.56 with a standard deviation of 0.022. The former estimate is unaffected by measurement error but has a rather large standard deviation. The latter estimate may be biased but has a small standard error. Neither estimate utilises information about the accuracy of the visual test contained in the cross-classification of D_T

Melodic Continuation

Melodic Beginning

	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12
-12																									
-11	○				○	○	○	●	○	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
-10	○				○	○	○	○	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	○
-9						○	○	○	○	●	●	●	●	●	●	●	●	●	○	●	●	●	●		○
-8				○	○	○	○	○	●	●	●	●	●	●	●	●	●	●	●	●	●	●	○		○
-7						○	○	●	○	○	●	●	●	●	●	●	●	●	●	●	●	●	○		●
-6																									
-5						●		○	●	●	●	●	●	●	●	●	●	●	●	●	●	●	○		
-4	○				○	○		●	●	●	●	●	●	●	●	●	●	●	○	●		○			
-3				○		○	○	●	●	●	●	●	●	●	●	●	●	●	○	●	○		○		
-2			○			○	○	○	○	●	●	●	●	●	●	●	●	●	○						
-1		○		○	○	○	○	●	●	●	●	●	●	●	●	●	●	●	○	○					
0	●					●	○	●	●	●	●	●	●	●	●	●	●	●		○		○			
1					○		○	○	○	●	○	●	●	●	●	●	●	●	○	○		○	○		
2			○		○	○		○	●	●	●	○	●	●	●	●	○	●							
3	○				○	●	○	●	●	●	●	●	●	●	●	●	●	●	○	○					
4		○		○			○	○	●	○	●	●	●	●	●	●	○	●				○			
5					○	○	○	●	●	●	●	●	●	●	●	●	●	●	○	●					
6																									
7	○			○	○	●	○	●	●	●	●	●	●	●	●	●	●	●		○	○				○
8	○	○	○	○	●	●	●	●	●	●	●	●	●	●	●	●	●	●	○	○					
9		○			○	●	●	●	●	●	●	●	●	●	●	●	○	●		○	○				
10	○	○		●	●	●	○		○	●	●	●	●	●	●	○	●	○							
11	●	●	○	●	○	●	●	●	○	●	●	●	●	●	●	●	○	○	○	○					
12																									

Figure 9: Fuzzy Empirical Expectancy Model Model. The "•"'s represent melodic continuations that are predicted to occur for each melodic beginning. The "∘"'s represent continuations for which the data is ambiguous. No predictions are made for the octave and tritone melodic beginnings (-12, -6, 6, and 12).

Visual and Pathology

	D_F	$\overline{D}_F$	
D_T	61	7	
$\overline{D}_T$	1	31	
			100

Visual Only

D_F	$\overline{D}_F$	
218	182	400

Table 7: Double Sampling: Jaundice Data

and D_F. What you would like to do is estimate $\mathrm{pr}(D_T)$ using *all* the data to hand, and ideally, estimate the accuracy of the visual test at the same time.

A straightforward maximum likelihood solution to this problem with attendant asymptotic standard error estimates was presented by Tenenbein (1970, 1972). Extensive generalisations of Tenenbein's work have been reported by Chen (1979, 1989), Ekholm and Palmgren (1987), Ekholm (1991), Espeland and Hui (1987), Espeland and Odoroff (1987), Lie *et al.* (1991a) and Nedelman (1988)—this list is by no means exhaustive. A Bayesian approach was presented by Geng and Asano (1989).

Here we present an approach to double sampling which is based on Bayesian graphical models. This allows us to account for model uncertainty, incorporate prior expert knowledge and tackle larger problems for which the conventional methods become unwieldy.

We begin by presenting in Figure 10 a trivial directed Bayesian graphical model for the jaundice example. As this is the only sensible model for this application, there is no model uncertainty. However, the graphical framework does facilitate the incorporation of prior knowledge through the elicitation of informative prior distributions for $\mathrm{pr}(D_T)$, $\mathrm{pr}(D_F \mid D_T)$ and $\mathrm{pr}(D_F \mid \overline{D}_T)$. Posterior distributions for quantities of interest are then derived via the Gibbs sampling method adopted in Section 4.

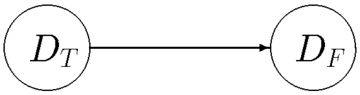


Figure 10: Double Sampling: Jaundice Data

For this fictituous example, Jeffreys prior distributions were used in place of informative prior distributions. The consequent "as if" posterior distribution for $\mathrm{pr}(D_T)$ is shown if Figure 11. The posterior mean and standard deviation for $\mathrm{pr}(D_T)$ are 0.63 and 0.026 respectively. This data was also analysed by Tenenbein (1970) and his corresponding estimates were 0.63 and 0.033. Point estimates for $\mathrm{pr}(D_F \mid \overline{D}_T)$ and $\mathrm{pr}(\overline{D}_F \mid D_T)$ are 0.034 and 0.131 for the Bayesian graphical model. Tenenbein's corresponding estimates are 0.025 and 0.129. The Bayesian graphical model estimates are quite insensitive to the choice of prior distribution.

Chen (1979), Chen *et al.* (1984) and Espeland and Odoroff (1985) introduce covariates, triple sampling and extra doubly sampled variables respectively into the above framework. The standard approach is to fit recursive systems of log linear models with maximum likelihood estimation via the EM algorithm. Ekholm and Palmgren (1987) adopt a more straightforward approach forming a single model with interpretable parameters. However, in each case, the analysis is characterized by tedious likelihood calculations and obscure derivations of asymptotic properties. The Bayesian graphical model approach by contrast, extends in a simple fashion to more complex models. Posterior distributions of many quantities of interest are easily derived.

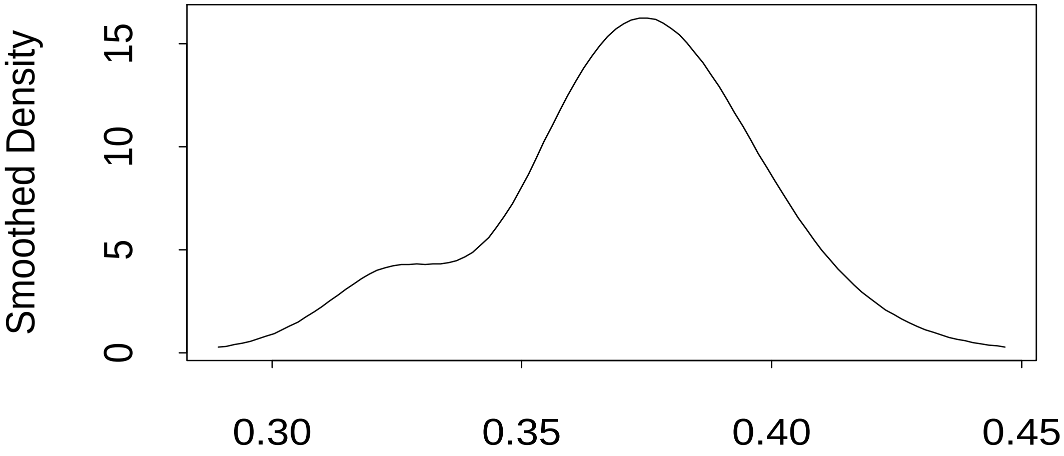


Figure 11: Posterior Distribution for pr(D_T)

5.2 Example : Down's Syndrome in Norway

We present the complete Bayesian graphical model approach to double sampling in the context of an example which was introduced by Lie *et al.* (1991a) and is further analyzed by York *et al.* (1992). Since 1970, epidemiological surveillance of congenital malformations has been carried out in Norway on the basis of data in the nationwide Medical Birth Registry (MBR). This data is collected at birth by the midwife or obstetrician and corresponds to the visual inspection in our fictitious jaundice example above. Because of growing concerns about incomplete ascertainment, a new notification system entitled "Melding om Fosterindiserte Aborter og Medfødte Misdannelser" (MIA) was introduced in 1985 in the county of Hordaland covering about 15% of all births in Norway. The MIA registration is based on prenatal diagnostics and pediatric follow-up including results from cytogenetic tests. However, unlike the fictituous jaundice example, the MIA registration *is* subject to error. Data concerning Down's syndrome collected between 1985 and 1988 is presented in Table 8. For further details, we refer the interested reader to Lie *et al.* (1991a,b).

Bayesian graphical models overcome two substantive difficulties with the analysis of this data presented by Lie *et al.* (1991a). First, although both of their models provide a reasonable fit to the data, Down's syndrome prevalence estimates and corresponding asymptotic standard errors are quite different under the two models. The Bayesian graphical model framework accounts for this model uncertainty. Second, Lie *et al.* (1991a) did not consider any covariates such as maternal age in their analysis. Because of the strong association with

Doubly Sampled Data

	R_1	$\overline{R}_1$	
R_2	8	9	
$\overline{R}_2$	13	17847	
			27877

Singly Sampled Data

R_1	$\overline{R}_1$	
233	188790	189023

Table 8: Down's syndrome data for 1985–1988 : R_1 represents case ascertainment through the national MBR registry and R_2 through the regional MIA registry.

maternal age, a complete study of the prevalence of Down's syndrome should include this covariate (Lie *et al.*, 1991b). However, the complexity of the existing analysis, in particular the calculation of asymptotic variances, suggests that such expansions would be difficult. Again, the Bayesian graphical model framework greatly facilitates both the incorporation of covariates.

The directed models we consider are subject to the constraint that links connecting error-free but possibly unobserved variables and error-prone observed variables are in the natural causal direction, i.e. from the error-free to the error-prone. A Markov chain Monte Carlo method was adopted for the analysis of the data of Table 8 augmented by maternal age (in six categories). Denoting by Δ, the prevalence of Down's syndrome, and by Y, the observed data, we want to compute $\text{pr}(\Delta \mid Y)$. To account for model uncertainty and integrate over Z, the missing data on the singly sampled cases, we re-express this as:

$$\text{pr}(\Delta \mid Y) = \sum \text{pr}(\Delta \mid M, Y, Z)\text{pr}(M, Z \mid Y)$$

where the summation is over all models, M, and all possible states of the missing data, Z. This can be numerically approximated by simulating a process $\{ Z(t), M(t) \}$ with stationary distribution $\text{pr}(Z, M \mid Y)$. A schematic version of the simulation method adopted is presented in Figure 12.

If necessary, simulating from $\text{pr}(M \mid Z, Y)$ can utilize a Metropolis step, as described in Section 2.

The results of a Bayesian graphical model analysis of the Down's syndrome data are given in Table 9 and Figure 13. In this analysis, all models were assumed equally likely *a priori* and informative prior distributions, based on historical data and expert knowledge were placed on the various probabilities. For details we refer the reader to York *et al.* (1992). The analysis assumes that there are no false positives, which is reasonable in this context. Models with a '*' on the R_1, R_2 link impose a special kind of dependence where it is assumed that the MIA registry, R_2, will find all cases missed by the national registry, R_1.

Except for the inclusion of the age covariate, the first two models in Table 9 correspond respectively to the two models examined by Lie *et al.* (1991a). Their first model produced a maximum likelihood estimate for $10^3 \times \text{pr}(S)$ of 2.02 with a standard deviation of 0.35, while

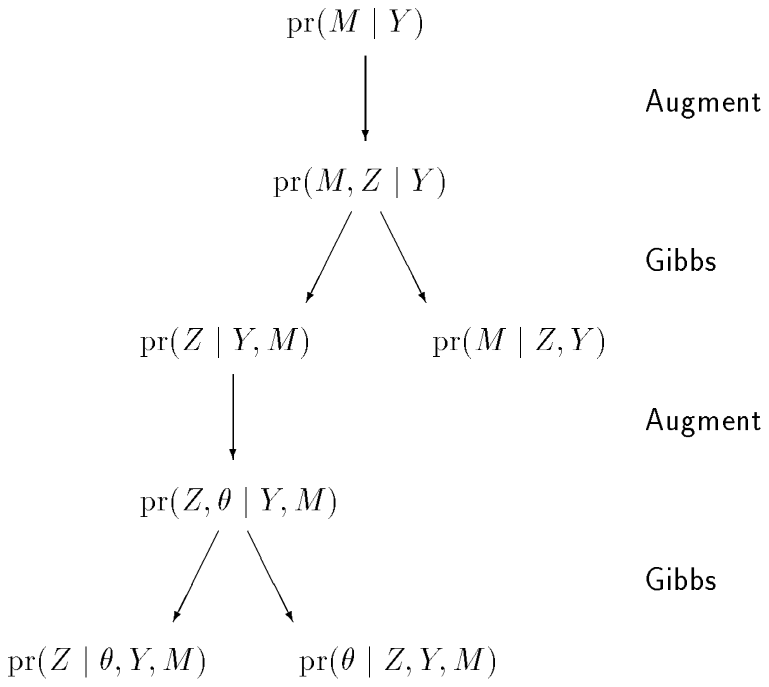

$\mathrm{pr}(M \mid Y)$
Augment
$\mathrm{pr}(M, Z \mid Y)$
Gibbs
$\mathrm{pr}(Z \mid Y, M)$
$\mathrm{pr}(M \mid Z, Y)$
Augment
$\mathrm{pr}(Z, \theta \mid Y, M)$
Gibbs
$\mathrm{pr}(Z \mid \theta, Y, M)$
$\mathrm{pr}(\theta \mid Z, Y, M)$

Figure 12: Markov Chain Monte Carlo Model Composition with missing data. In order to generate a process with the stationary distribution given at the top of the tree, we simulate iteratively from the distributions at the leaves of the tree.

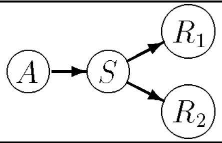
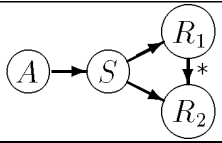
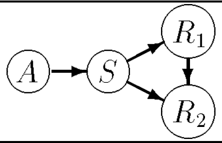
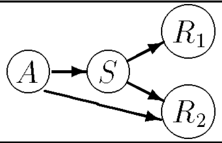
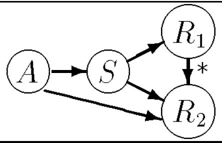
	Post.	$10^3 \times \mathrm{pr}(S)$			$\mathrm{pr}(\overline{R}_1 \mid S)$		$\mathrm{pr}(\overline{R}_2 \mid S)$	
Model	Prob.	Mode	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
	0.282	1.81	1.92	0.292	0.376	0.085	0.555	0.092
	0.385	1.49	1.51	0.129	0.223	0.053	0.470	0.083
	0.269	1.60	1.70	0.252	0.312	0.088	0.513	0.089
	0.030	1.71	1.78	0.226	0.333	0.076	0.518	0.090
	0.016	1.50	1.52	0.129	0.226	0.054	0.517	0.080
Model Averaging	—	1.54	1.69	0.289	0.292	0.099	0.508	0.095

Table 9: Features of the posterior for Down's syndrome prevalence and the error probabilities of the two registries. Prevalence is given as the rate per thousand. Only models with posterior probability larger than 0.01 are listed; all models are included in the model averaging results.

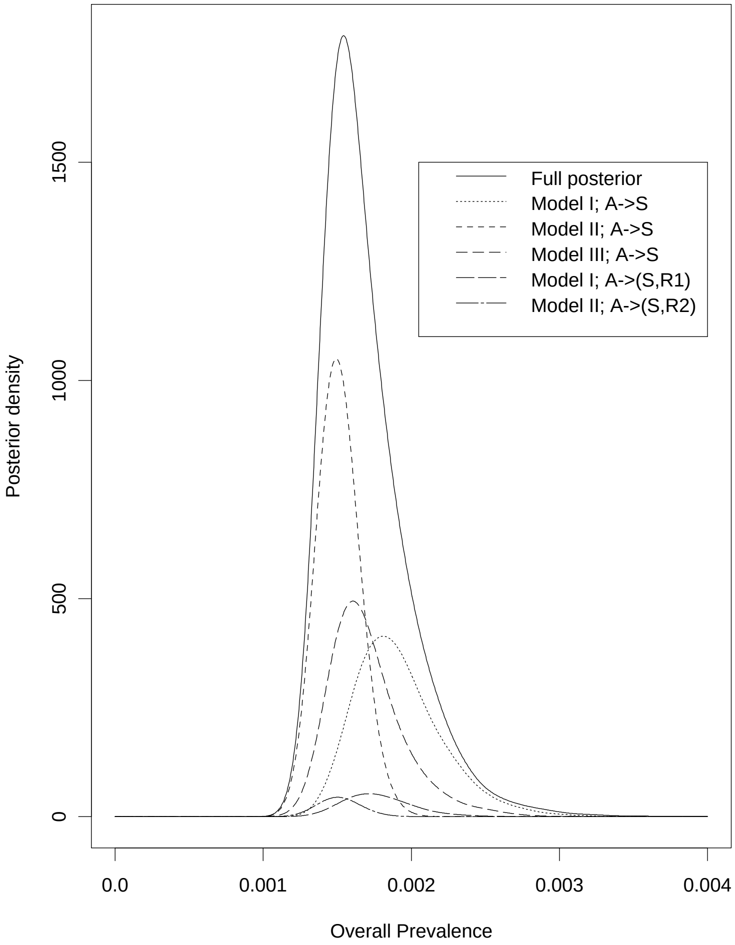


Figure 13: Overall posterior for Down's syndrome rate per 1000 when the mother's age is included as a covariate, along with the posterior for each individual model scaled according to its posterior probability.

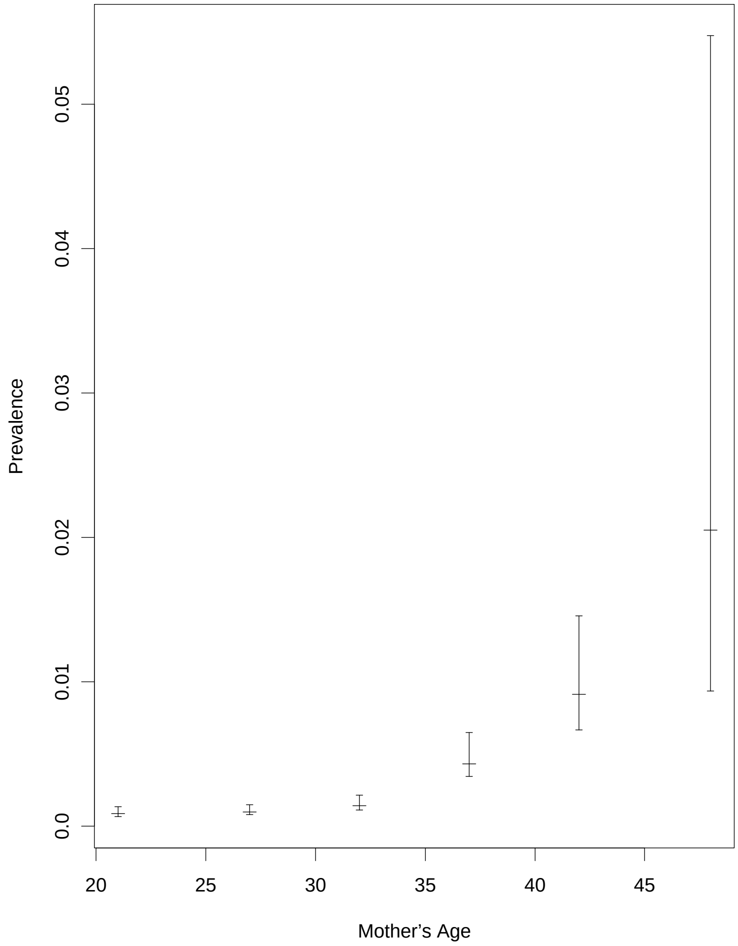


Figure 14: Mode and 5th and 95th percentiles of the posterior for Down's syndrome prevalence by age of mother, averaged across all models.

their second model gives 1.49 and 0.13. Our analysis accounts for the this model uncertainty, averaging over all the models. Furthermore, incorporation of the maternal age substantially improves model fit and allows for age-specific reporting, such as in Figure 14.

5.3 Why Bayesian Graphical Models?

Bayesian graphical models extend the reach of multiply sampled data analysis into heretofore intractable areas. Models of considerable complexity can be considered and posterior distributions for a variety of quantities of interest derived. Expert knowledge can realistically be incorporated and model uncertainty can be accounted for.

One note of caution: the Markov chain Monte Carlo method outlined here may run into some practical difficulties in the analysis of *very* large datasets. The essential problem is that the missing data conveys considerable information about the best models. Consequently, their joint distribution, $\mathrm{pr}(M, Z \mid Y)$ can be highly multimodal. We are currently investigating possible solutions to this problem–see also Besag and Green (1993) and Lin (1992).

6 Data Quality: Predicting Errors in Databases

6.1 Introduction

A recent article by Strayhorn (1990) introduced an important class of problems in data quality management. The techniques developed potentially have wide application in quality control or indeed in any environment where flawed items must be detected and counted. Strayhorn was motivated specifically by the quality control of research data. He points out that while large numbers of journal pages are devoted to the quantification and control of measurement error, possible errors in data are rarely mentioned (see Feigl *et al.*, 1982, for a notable exception).

Strayhorn (1990) presented two methods for estimating error rates in research data: the duplicate performance method and the known errors method. However, his analysis was heavily criticized by West and Winkler (1991), hereafter referred to as WW, who present Bayesian analyses of the two methods. Madigan *et al.* (1993b) introduce a third method, the duplicate checking method, and show how Bayesian Graphical models provide for a simple and extensible analysis of all three methods.

Here we briefly describe these Bayesian graphical models and the possibilities they present.

6.2 Duplicate Performance Method

Suppose that a large number, N, of paper-based medical records must be entered into a computer database, and further suppose that two data entry personnel, α and ω are available to carry out this task. The idea is that both independently key in the data and then the resulting computer files are compared item by item by a method assumed to be error free. Where there is disagreement, the original paper record is consulted and the disagreement settled. Let d be the total number of disagreements found in this way, $d = x_\alpha + x_\omega$, where x_j is the number of errors attributable to j, $j = \alpha, \omega$. The only errors remaining are the subset of the $N - d$ records where *both* α and ω were in error. The intuition is that if the ratio of disagreements to total items, $\frac{d}{N}$, is low, the individual error rates of α and ω are low, and the probability of joint errors is lower still.

Because α and ω carry out their tasks independently a trivial Bayesian graphical model for this situation has two unconnected nodes A_α and A_ω, where A_j is a binary random variable indicating whether j entered a particular record correctly or not, $j = \alpha, \omega$. WW suggest that in practice d/N will typically be small so that agreement between α and ω will occur for most records. For binary records, this sort of data will often be equally consistent with both typists being almost always correct or both being almost always incorrect. Consequently, uniform $[0, 1]$ priors on $\text{pr}(A_\alpha)$ and $\text{pr}(A_\omega)$ will result in heavily bimodal posterior distributions. To counteract this problem, WW put prior distributions on $\text{pr}(A_\alpha)$ and $\text{pr}(A_\omega)$ which only include values larger than 0.5 in their support. This takes them outside the class of conjugate priors however. This bimodality problem can also be avoided by using informative priors which are centered on a value greater than 0.5, thereby assuming *a priori* that the typists are more likely to enter data correctly than not. This approach retains conjugacy which proves very useful when performing the calculations. Furthermore prior distributions which are truncated at 0.5, especially the uniform prior on $[0.5, 1]$, will typically provide a poor model for prior expert knowledge.

	A_ω	$\overline{A}_\omega$	
A_α	z	x_ω	
$\overline{A}_\alpha$	x_α	$N - x_\alpha - x_\omega - z$	
			N

Table 10: Duplicate Performance Method Table

The framework for this method may be represented as a 2×2 table—see Table 10. Here z represents the number of records where α and ω are both correct. Then we have:

$$\begin{aligned} \text{pr}(z \mid x_\alpha, x_\omega, N) &\propto \text{pr}(z, x_\alpha, x_\omega \mid N) \\ &= \int_\theta \text{pr}(D \mid N, \theta)\text{pr}(\theta) d\theta \end{aligned}$$

where D represents complete data and θ is the vector parameter for the cell probabilities.

In Table 11 we present results for some of the hypothetical datasets considered by Strayhorn and WW assuming "informative" prior beta(1,3) distributions for $\text{pr}(A_\alpha)$ and $\text{pr}(A_\omega)$. This assigns both quantities a prior mean of 0.75 and standard deviation of 0.19. For each dataset we show the probabilities of various undetected error counts. Also provided is the probability assigned to the event that all the events on which there is agreement are in error–this is to demonstrate that the bimodality problem is adequately addressed through the use of reasonable informative priors. The probability of zero undetected errors is included from the WW analysis for comparison purposes.

				WW	$\mathrm{pr}(z \mid x_\alpha, x_\omega)$							
n	x_α	x_ω	$\mathrm{pr}(z > 0 \mid x_\alpha, x_\omega)$	z=0	z=0	z=1	z=2	z=3	z=4	z=5	z=6	z=max
20	2	3	0.46	0.27	0.54	0.26	0.10	0.04	0.02	0.01	0.01	0.00
20	1	1	0.17	0.59	0.83	0.14	0.03	0.01	0.00	0.00	0.00	0.00
20	1	0	0.09	0.71	0.91	0.08	0.01	0.00	0.00	0.00	0.00	0.00
5000	50	50	0.41	0.58	0.59	0.31	0.08	0.02	0.00	0.00	0.00	0.00
5000	25	25	0.13	0.86	0.87	0.12	0.01	0.00	0.00	0.00	0.00	0.00
5000	5	5	0.01	0.99	0.99	0.01	0.00	0.00	0.00	0.00	0.00	0.00
5000	2	3	0.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 11: Duplicate Performance Method: Hypothetical Data and Predictive Probabilities for Undetected Errors, Independent Be(1,3) Priors for $\mathrm{pr}(A_\alpha)$ and $\mathrm{pr}(A_\omega)$. n is the total number of records, x_α and x_ω are the number of errors attributable to each of the two checkers and z is the number of undetected errors.

6.3 Duplicate Checking Method

A somewhat different approach was alluded to but not analyzed by WW, and we refer to this as the duplicate checking method. Here we assume that the database *already exists* and the task of our two friends α and ω is to independently check each record in the database. WW make an important assumption that error free records are classified correctly although our analysis does not require this assumption. Thus the method may be represented as in Figure 15 where D_i now indicates whether or not i detected an error and X is a binary random variable indicating whether or not the record in the database is correct. The key point is that we have an extra piece of information here, namely the number of records for which both α and ω detect errors. This is similar to several of the previous examples and again informative prior distributions can be readily elicited in terms of well-understood quantities. We refer the reader to Madigan *et al.* (1993b) for numerical examples.

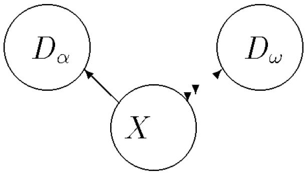


Figure 15: Duplicate Checking Method

6.4 Known Errors Method

The known errors method is described by Strayhorn (1990) as follows: "In this method, a member of the research staff completes the data operation in question. The data are

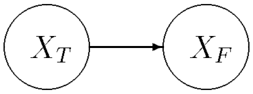


Figure 16: Known Errors Method

then presented to a second person, for example, the supervisor of the staff member, who introduces a certain number of 'known errors' into the data set. The locations and forms of these errors are recorded elsewhere. Then the data set together with known and unknown errors, is given to another staff member, who checks the data set."

WW provide two elegant analyses of this method. Our purpose here is to point out that the known errors method is a special case of double sampling. Here we have a simple Bayesian graphical model with two nodes, X_T, representing the true state of the record, and X_F, a "fallible" version representing what the checker has recorded. For the original data we only have observations on X_F while for the known errors, both nodes are recorded. To be consistent with the analysis of WW, uniform prior distributions were used. Note that the *known* values of X_T are not used when updating the distribution of $\mathrm{pr}(X_T)$.

The results presented in Madigan *et al.* (1993b) are essentially identical to those of WW.

6.5 Why Bayesian graphical models?

We have outlined how directed Bayesian graphical models provide for a straightforward analysis of three database error checking methodologies. In each case informative prior distributions can easily be specified in terms of readily understood quantities and modeling assumptions are transparent. The calculations in each case are straightforward, providing outputs which are much easier to interpret than Strayhorn's confidence intervals.

However, the real strength of the Bayesian graphical modeling approach for these problems is that it can be generalised in a simple fashion. In particular, the generalizations suggested by WW and Madigan *et al.* (1993b), can easily be incorporated. These include relaxation of the no-false-positive assumptions, varying error rate probabilities according to some characteristic of the data records, adding additional checkers, mixing duplicate and known errors methods, and sampling only a portion of the database.

7 Discussion

We have attempted to show that Bayesian graphical models represent a powerful unified framework for a wide variety of discrete data problems. Modeling assumptions are entirely transparent and computations are simple to program. Expert knowledge can easily be incorporated and model uncertainty accounted for.

The methods we discuss can readily be extended in two particular directions. First, graphical Gaussian models could be included. These were introduced as covariance selection models by Dempster (1972) and are discussed in Whittaker (1990). The variables being modeled in a graphical Gaussian model have a multivariate normal distribution. Conditional independencies, which correspond to zeroes in the inverse variance, are represented by an undirected graph. The Bayesian framework for these models has been developed by Dawid and Lauritzen (1993). Recent extensions to this model class described by Cox and Wermuth (1993) are of considerable interest in this context.

Second, the graphs we consider here are either undirected or fully directed. The methods could be extended to include chain independence graphs, also called block recursive graphs by Lauritzen and Wermuth (1989). These graphs may both directed and undirected links and provide support for a richer class of models.

A very valuable development would be to include the mixed discrete/continuous models of Wermuth and Lauritzen (1990) and Edwards (1990).

Appendix I: Graph Theoretic Terminology

The terminology we use is largely adapted from Lauritzen *et al.* (1990).

A *graph* is a pair $G = (V, E)$ where V is a finite set of vertices and the set of edges, E, is a subset of $V \times V$ of ordered pairs of distinct vertices. Edges $(\alpha, \beta) \in E$ with both (α, β) and (β, α) in E are called *undirected*, whereas an edge (α, β) with its opposite (β, α) not in E is called *directed.*

If the graph has only undirected edges it is *undirected* and if all the edges are directed, the graph is said to be *directed.* Our graphs are either directed or undirected.

If $A \subseteq V$ is a subset of the vertex set, it induces a subgraph $G_A = (A, E_A)$, where the edge set $E_A = E \cap (A \times A)$ is obtained from G by keeping edges with both endpoints in A.

A graph is *complete* if all vertices are joined by an edge. A subset is *complete* if it induces a complete subgraph. A complete subset that is maximal with respect to inclusion is called a *clique.*

In a directed graph, if $(\alpha, \beta) \in E$, α is said to be a *parent* of β and β a *child* of α. The set of parents of β is denoted by $\mathrm{pa}(\beta)$ and the set of children by $\mathrm{ch}(\beta)$.

In an undirected graph, if $(\alpha, \beta) \in E$, α and β are said to be *adjacent* or *neighbours.* The *boundary*, $\mathrm{bd}(A)$, of a subset A of vertices is the set of vertices in $V \backslash A$ that are neighbours to vertices in A. The *closure* of A is $\mathrm{cl}(A) = A \cup \mathrm{bd}(A)$.

A *path* of length n from α to β is a sequence

$\alpha = \alpha_0, \ldots, \alpha_n = \beta$ of distinct vertices such that $(\alpha_{i-1}, \alpha_i) \in E$ for all $i = 1, \ldots, n$. If there is a path from α to β we say that α *leads to* β and write $\alpha \mapsto \beta$. The *descendants* $\mathrm{de}(\alpha)$ of α are all the vertices β such that α *leads to* β. The *nondescendants* are $\mathrm{nd}(\alpha) = V \backslash (\mathrm{de}(\alpha) \cup \{\alpha\})$. The vertices α that lead to β are called the *ancestors* of β, denoted by $\mathrm{an}(\beta)$.

A subset $A \subseteq V$ is an *ancestral* set if it contains all its own ancestors, i.e if $\mathrm{an}(\alpha) \subseteq A$ for all $\alpha \in A$.

A *chain* of length n from α to β is a sequence $\alpha = \alpha_0, \ldots, \alpha_n = \beta$ of distinct vertices such that $(\alpha_{i-1}, \alpha_i) \in E$ *or* $(\alpha_i, \alpha_{i-1}) \in E$ for all $i = 1, \ldots, n$.

A subset S is said to *separate* A *from* B if all chains from vertices $\alpha \in A$ to $\beta \in B$ intersect S.

A *cycle* is a path with the modification that $\alpha = \beta$, i.e. it begins and ends at the same point. A directed graph is *acyclic* is it contains no cycles. An undirected graph is *chordal* if it contains no cycles of length ≥ 4 without a chord (i.e two non-consecutive vertices that are neighbours).

An ordering of the cliques of an undirected graph, say $(C_1, \ldots, C_n)$ is said to be *perfect* if the nodes of each clique C_i also contained in previous cliques $(C_1, \ldots, C_{i-1})$ are all members of *one* previous clique. These sets $S_i = C_i \cap (\cup_{j=1}^{i-1} C_j)$ are called *clique separators*. An undirected graph admits a perfect ordering of its cliques if and only if it is chordal.

For a directed acyclic graph $G^<$, we define its *moral graph*, G^m as the undirected graph with the same vertex set but with α and β adjacent in G^m if and only if either $(\alpha, \beta) \in E$ or $(\beta, \alpha) \in E$ or there exists a γ such that $(\alpha, \gamma) \in E$ and $(\beta, \gamma) \in E$. In other words the moral graph is obtained from the original graph by 'marrying parents' with a common child and the then dropping the directions on the edges.

References

Besag, J. and Green, P.J. (1993) Spatial statistics and Bayesian computation (with discussion). *Journal of the Royal Statistical Society (Series B)*, **55**,25–38.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975) *Discrete Multivariate Analysis* Cambridge, Mass.: MIT Press.

Bissell, A.D. (1921) The role of expectation in music. Unpublished doctoral dissertation, Yale University.

Breiman, L. (1968) *Probability.* Addison-Wesley, Reading.

Breslow, N. (1991) Biostatistics and Bayes. *Statistical Science*, **5**,269–298.

Carlsen, J.C. (1981) Some factors which influence melodic expectancy. *Psychomusicology*, **1**,12–29.

Carlsen, J.C. and Unyk, A.M. (1987) The influence of expectancy on melodic perception. *Psychomusicology*, **7**,3–23.

Carlsen, J.C., Bradshaw, D.H., Madigan, D., and Unyk, A.M. (1992) Music Expectancy and its measurement (Abstract). *Second International Conference on Music Perception and Cognition*, Los Angeles (also submitted for publication).

Charniak, E. (1991) Bayesian networks without tears. *AI Magazine*, **??**,50–63.

Chen, T.T. (1979) Log-linear models for categorical data with misclassification and double sampling. *Journal of the American Statistical Association*, **74**,481–488.

Chen, T.T., Hochberg, Y., and Tenenbein, A. (1984) On triple sampling schemes for categorical data analysis with misclassification errors. *Journal of Statistical Planning and Inference*, **9**,177–184.

Chen, T.T. (1989) A review of methods for misclassified categorical data in epidemiology. *Statistics in Medicine*, **8**,1095–1106.

Chung, K.L. (1967) *Markov Chains with Stationary Transition Probabilities* (2nd ed). Springer–Verlag, Berlin.

Cooper, G.F. and Herskovits, E. (1992) A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, **9**,309–347.

Cox, D.R. and Wermuth, N. (1993) Linear dependencies represented by chain graphs. *Statistical Science*, To appear.

Dawid, A.P. and Lauritzen, S.L. (1993) Hyper-Markov laws in the statistical analysis of decomposable graphical models *Annals of Statistics*, To appear.

Dawid, A.P. (1992) Applications of a general propagation algorithm for probabilistic expert systems. *Statistics and Computing*, **2**,25–36.

Dempster, A.P. (1972) Covariance selection. *Biometrics*, **28**,157–175.

Draper, D., Hodges, J.S., Leamer, E.E., Morris, C.N., and Rubin, D.B. (1987) A research agenda for assessment and propagation of model uncertainty. Rand Note N-2683-RC, The RAND Corporation, Santa Monica, California.

Draper, D. (1993) Assessment and propagation of model uncertainty. Submitted for publication.

Edwards, D. (1990) Hierarchical mixed interaction models. *Journal of the Royal Statistical Society (Series B)*, **52**,3–20.

Ekholm, A. (1991) Algorithms versus models for analyzing data that contain misclassification errors. *Biometrics*, **47**,1171–1182.

Ekholm, A. and Palmgren, J. (1987) Correction for misclassification using doubly sampled data. *Journal of Official Statistics*, **3**,419–429.

El–Khorazaty, M.N., Imrey, P.B., Koch, G.G., and Bradley, H. (1977) Estimating the total number of events with data from multiple–record systems : a review of methodological strategies *International Statistical Review*, **45**, 129–157.29–157.

Espeland, M.A. and Odoroff, C.L. (1985) Log-linear models for doubly sampled categorical data fitted by the EM algorithm. *Journal of the American Statistical Association*, **80**,663–670.

Espeland, M.A. and Hui, S.L. (1987) A general approach to analyzing epidemiologic data that contain misclassification errors. *Biometrics*, **43**,1001–1012.

Feigl, P., Polissar, L., Lane, W., and Guinee, V. (1982) Reliability of basic cancer patient data. *Statistics in Medicine*, **1**, 191–204.

Fienberg, S.E. (1972) The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika*, **59**, 591–603.

Frydenberg, M. and Lauritzen, S.L. (1989) Decomposition of maximum likelihood in mixed graphical interaction models. *Biometrika*, **76**,539–555.

Frydenberg, M. (1990) The chain graph Markov property *Scandinavian Journal of Statistics*, **17**,333–353.

Geng, Z. and Asano, C. (1989) Bayesian estimation methods for categorical data with misclassifications. *Communications in Statistics: Theory and Methods*, **18**,2935–2954.

Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**,97–109.

Hodges, J.S. (1987) Uncertainty, policy analysis and statistics. *Statistical Science*, **2**,259–291.

Hook, E.B., Albright, S.G., and Cross, P.K. (1980) Use of Bernoulli census and log–linear methods for estimating the prevalence of Spina Bifida in livebirths and the completeness of vital record reports in New York state. *American Journal of Epidemiology*, **112**, 750–758.

Kiely, E.A., Madigan, D., Ryan, P.C., and Butler, M.R. (1990) Ultrasonic imaging for extracorporeal shockwave lithotripsy: analysis of factors in successful treatment. *British Journal of Urology*, **66**,127–131.

Kiiveri, H., Speed, T.P., and Carlin, J.B. (1984) Recursive Causal Models. *Journal of the Australian Mathematical Society (Series A)*, **36**,30–52.

Kornfeld, A. (1991) Causal networks: Clarifying uncertainty. *AI Expert*, November, 42–49.

Lange, N. (1992). Graphs and stochastic relaxation for hierarchical bayes modeling. *Statistics in Medicine*, **11**, 2001–2016.

Lauritzen, S.L. (1992) Propagation of probabilities, means, and variances in mixed graphical association models. *Journal of the American Statistical Association*, **87**,1098–1108.

Lauritzen, S.L., Dawid, A.P., Larsen, B.N., and Leimer, H-G. (1990) Independence properties of directed markov fields, *Networks*, **20**:491–505.

Lauritzen, S.L. and Spiegelhalter, D.J. (1988) Local computations with probabilities on graphical structures and their application to expert systems (with discussion.) *Journal of the Royal Statistical Society (Series B)*, **50**,157–224.

Lauritzen, S.L. and Wermuth, N. (1989) Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, **17**,31–57.

Lie, R.T., Heuch, I., and Irgens, L.M. (1991a) Estimation of the proportion of congenital malformations using double registration schemes. Submitted for publication.

Lie, R.T., Heuch, I., and Irgens, L.M. (1991b) A temporary increase of Down syndrome among births of young mothers in Norway: An effect of risk unrelated to maternal age?. *Genetic Epidemiology*, **8**,217–230.

Lin, S. (1992) On the performance of Markov chain Monte Carlo methods on pedigree data and a new algorithm. *Technical Report 231*, Department of Statistics, University of Washington.

Madigan, D., Carlsen, J.C., and Bradshaw, D.H. (1992) Development of a data-based expectancy model (Abstract). *Second International Conference on Music Perception and Cognition*, Los Angeles.

Madigan, D. and Mosurski, K.R. (1990) An extension of the results of Asmussen and Edwards on collapsibility in contingency tables. *Biometrika*, **77**,315–319.

Madigan, D. and Mosurski, K.R. (1991) Explanation in belief networks. Submitted for publication.

Madigan, D. and Raftery, A.E. (1991) Model selection and accounting for model uncertainty in graphical models using Occam's window. *Technical Report 213*, Department of Statistics, University of Washington.

Madigan, D., Raftery, A.E., York, J.C., Bradshaw, J.M., and Almond, R.G. (1993a) Strategies for graphical model selection. *Proceedings of the Fourth International Workshop on Artificial Intelligence and Statistics*, Florida, to appear.

Madigan, D., York, J.C., Bradshaw, J.M., and Almond, R.G. (1993b) Bayesian graphical models for predicting errors in databases. *Proceedings of the Fourth International Workshop on Artificial Intelligence and Statistics*, Florida, to appear.

Merkins, G.E. and Anderson, S.H. (1988) Estimation of small–mammal population size. *Ecology*, **69**, 1952–1959.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.

Miller, A.J. (1984) Selection of subsets of regression variables (with Discussion). *Journal of the Royal Statistical Society (Series A)*, **147**, 389–425.

Moulton, B.R. (1991) A Bayesian approach to regression selection and estimation with application to a price index for radio services. *Journal of Econometrics*, to appear.

Narmour, E. (1990) *The Analysis and Cognition of Basic Melodic Structures: The Implication-Realization Model,* Volume 1. Chicago: University of Chicago Press.

Neapolitan, R.E. (1990) *Probabilistic Reasoning in Expert Systems* New York: Wiley.

Nedelman, J. (1988) The prevalence of malaria in Garki, Nigeria: double sampling with a fallible expert. *Biometrics*, **44**,635–655.

Otis, D.L., Burnham, K.P., White, G.C., and Anderson, D.R. (1978) Statistical inference for capture data on closed animal populations. *Wildlife Monographs*, **62**, 1–135.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers Inc., San Mateo, California.

Pollock, K.H. and Otto, M.C. (1983) Robust estimation of population size in closed animal populations from capture–recapture experiments. *Biometrics*, **39**, 1035–1049.

Raftery, A.E. (1988a) Inference for the binomial N parameter : a hierarchical Bayes approach. *Biometrika*, **75**, 223–228.

Raftery, A.E. (1988b) Approximate Bayes factors for generalised linear models. *Technical Report 121*, Department of Statistics, University of Washington.

Raftery, A.E. (1992) Bayesian model selection in structural equation models. In *Testing Structural Equation Models* (eds. K.A. Bollen and J.S. Long), Beverly Hills: Sage.

Regal, R.R. and Hook, E.B. (1991) The effects of model selection on confidence intervals for the size of a closed population. *Statistics in Medicine* **10**, 717–721.

Rodrigues, J., Bolfarine, H., and Leite, J.G. (1988) A Bayesian analysis in closed animal populations from capture recapture experiments with trap response. *Communications in Statistics – Simulation*, **17**, 407–430.

Smith, A.F.M. and Roberts, G.O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *Journal of the Royal Statistical Society (Series B)*, **55**,3–23.

Spiegelhalter, D.J. (1986) Probabilistic reasoning in expert systems. in *Uncertainty in Artificial Intelligence*, (eds. L.N. Kanal and J. Lemmer). North-Holland.

Spiegelhalter, D.J. and Lauritzen, S.L. (1990) Sequential updating of conditional probabilities on directed graphical structures. *Networks*, **20**,579–605.

Spiegelhalter, D.J., Dawid, A.P., Lauritzen, S.L., and Cowell, R.G. (1993) Bayesian analysis in expert systems. *Statistical Science.* To appear.

Strayhorn, J.M. (1990) Errors remaining in a data set: techniques for quality control. *The American Statistician*, **44**, 14–18.

Tanner, M.A. and Wong, W.H. (1987) The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, **82**,528–550.

Tenenbein, A. (1970) A double sampling scheme for estimating from binomial data with misclassification. *Journal of the American Statistical Association*, **65**,1350–1361.

Tenenbein, A. (1972) A double sampling scheme for estimating from misclassified multinomial data with application to sampling inspection. *Technometrics*, **14**,187–202.

Titterington, D.M. (1985) Common structure of smoothing techniques is statistics. *International Statistics Review*, **53**, 141–170.

Wermuth, N. and Lauritzen, S.L. (1990) On substantive research hypotheses, conditional independence graphs and graphical chain models. *Journal of the Royal Statistical Society (Series B)*, **52**,21–50.

West, M. and Winkler, R.L. (1991) Data base error trapping and prediction. *Journal of the American Statistical Association*, **86**,987–996.

Whittaker, J. (1990) *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.

Wolter, K.M. (1986) Some coverage error models for census data. *JASA*, **81**, 338–346.

Wright, S. (1921) Correlation and causation. *Journal of Agricultural Research*, **20**, 557–585.

York, J. and Madigan, D. (1992) Bayesian methods for estimating the size of a closed population. Technical Report 234, Department of Statistics, University of Washington. Submitted for publication

York, J., Madigan, D., Heuch, I., and Lie, R.T. (1992) Estimating a proportion of birth defects by double sampling: A Bayesian approach incorporating covariates and model uncertainty. Submitted for publication.