Bayesian Data Analysis for Data Mining

David Madigan
Rutgers University
madigan@stat.rutgers.edu

Greg Ridgeway
RAND
gregr@rand.org

June 13, 2002

1 Introduction

The Bayesian approach to data analysis computes conditional probability distributions of quantities of interest (such as future observables) given the observed data. Bayesian analyses usually begin with a *full probability model* - a joint probability distribution for all the observable and unobservable quantities under study - and then use Bayes' theorem (Bayes, 1763) to compute the requisite conditional probability distributions (called *posterior distributions*). The theorem itself is innocuous enough. In its simplest form, if Q denotes a quantity of interest and D denotes data, the theorem states:

$$p(Q|D) = p(D|Q) \times p(Q)/p(D).$$

This theorem prescribes the basis for statistical learning in the probabilistic framework. With $p(Q)$ regarded as a probabilistic statement of prior knowledge about Q before obtaining the data D, $p(Q|D)$ becomes a revised probabilistic statement of our knowledge about Q in the light of the data (Bernardo and Smith, 1994, p.2). The marginal likelihood of the data, $p(D)$, serves as normalizing constant.

Computing is the big issue confronting a data miner working in the Bayesian framework. The computations required by Bayes' theorem can be demanding, especially

with large datasets. In fact, widespread application of Bayesian data analysis methods has only occurred in the last decade or so, having had to wait for computing power as well as breakthroughs in simulation technology. Barriers still exist for truly large-scale applications.

The primary advantages of the Bayesian approach are its conceptual simplicity and the commonsense interpretation of Bayesian outputs. The ability to incorporate prior knowledge can also be a boon. Many data mining applications provide copious data, but for models with thousands if not millions of dimensions or parameters, a limited amount of prior knowledge, often in the form of prior exchangeability information, can sharpen inferences considerably. Perhaps more commonly though, the available data simply swamps whatever prior knowledge is available, and the precise specification of the prior becomes irrelevant.

Section 2 of this chapter uses a series of examples to introduce the basic elements of the Bayesian approach. Sections 4 and 5 describe strategies for Bayesian computation and model building respectively. The remaining Sections discuss some specific applications and describe currently available software.

2 Fundamentals of Bayesian Inference

2.1 A Simple Example

Consider estimating the "sex ratio," that is, the proportion of births that are female, in a specific population of human births. Demographers have studied sex ratios for centuries and variations across sub-populations continue to attract research attention. In 1781 the illustrious French scientist Pierre Simon Laplace presented a Bayesian analysis of the sex ratio (Stigler, 1990). He used data concerning 493,472 Parisian births between 1745 and 1770. Let $n = 493,472$ and $y_1, \ldots, y_n$ denote the sex associated with each of the births, $y_i = 1$ if the ith birth is female $y_i = 0$ is the ith birth is male. For the Parisian data $\sum y_i = 241,945$ and we denote this by y. Denote by θ the probability that a given birth is female. Assuming that the births

represent independent and identically distributed Bernoulli trials, we can compute the posterior distribution as:

$$p(\theta|y_1,\ldots,y_n) \propto \theta^y(1-\theta)^{n-y}p(\theta), \theta \in [0,1].$$

Ultimately we will want to use this posterior distribution to compute probabilities such $\Pr(\theta \geq \frac{1}{2})$ or $\Pr(0.49 \leq \theta \leq 0.50)$. However, to do this we will need to compute the normalizing constant on the right hand side, that is,

$$\int \theta^y(1-\theta)^{n-y}p(\theta)d\theta,$$

and we are getting a first glimpse as to why computation looms large in Bayesian analysis. Since the integral here is one-dimensional, it yields a straighforward numerical approximation irrespective of the precise choice of $p(\theta)$ (see Section 4). Laplace circumvented the integration by using a uniform prior distribution for θ (i.e., $p(\theta) = 1, \theta \in [0,1]$) leading to:

$$p(\theta|y_1,\ldots,y_n) \propto \theta^{241945}(1-\theta)^{251527}, \theta \in [0,1].$$

This is the density of the so-called beta distribution and, in this case, the requisite normalizing constant is available in closed form:

$$p(\theta|y_1,\ldots,y_n) = \frac{\Gamma(493474)}{\Gamma(241946)\Gamma(251528)}\theta^{241945}(1-\theta)^{251527}, \theta \in [0,1].$$

The posterior mean and standard deviation are 0.490291 and 0.000712 respectively. Could the true probability be as big as a half? That was the scientific question that Laplace addressed and the posterior density above yields:

$$\Pr(\theta \geq \frac{1}{2}) \approx 1.521 \times 10^{-42}.$$

Laplace concluded that it is "morally certain" that the true probability is indeed less than a half. Note that this analysis, unlike a classical p-value, provides a direct probabilistic answer to the primary question of interest.

What about prediction? Denote by y_{n+1} the sex associated with the next birth. We compute the predictive density as follows:

$$p(y_{n+1}|y_1,\ldots,y_n) = \int p(y_{n+1}|\theta)p(\theta|y_1,\ldots,y_n)d\theta.$$

Note that we have used here the fact the θ renders y_{n+1} independent of $y_1, \ldots, y_n$. Again, since we are dealing with a one-dimensional parameter, this integral is manageable. With Laplace's particular choice of prior, the predictive distribution takes an especially simple form:

$$\Pr(y_{n+1} = 1 | y_1, \ldots, y_n) = \frac{y+1}{n+2} = 0.490291.$$

Laplace's choice of the uniform prior led to closed-form expressions for the posterior and the predictive distributions. He could more generally have chosen any beta distribution as the prior distribution for θ and still have closed-form outputs. The general form of the beta density for a θ is:

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \theta \in [0,1].$$

We refer to this as a beta(α, β) distribution where α and β are *hyperparameters*. This leads to a beta$(\alpha + y, \beta + n - y)$ posterior distribution. For Bernoulli/binomial data the beta distribution is said to be the *conjugate prior distribution*. Conjugate priors stay in the same distributional family as they go from prior to posterior. Convenient conjugate priors are available for many standard distributions and, so long as they are flexible enough to accurately represent whatever prior knowledge is available, it often makes sense to use a conjugate prior. The BUGS manual (Spiegelhalter *et al.*, 1999) provides a list of distributions and their associated conjugate priors. Gelman *et al.* (1995) also analyze the example just described and provide a comprehensive introduction to Bayesian methods in general.

2.2 A More Complicated Example

The use of graphs to represent statistical models has a rich history dating back at least to the 1920's. Recently, graphical Markov models have emerged as a important class of models and have impacted fields such as data mining, causal analysis, and statistical learning. A graphical Markov model is a multivariate probabilistic model that uses a graph to represent a set of conditional independences. The vertices of the graph represent the random variables of the model and the edges encode the

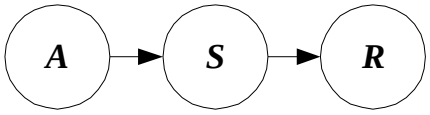


Figure 1: Down's Syndrome: An Acyclic Directed Graphical Markov Model

conditional independences. In general, each missing edge corresponds to a conditional independence. Graphs with different types of edges–directed, undirected, or both–lead to different classes of probabilistic models. In what follows we will only consider acyclic directed models, also known as *Bayesian Networks* [1] (see, for example, Pearl, 1988).

Spiegelhalter and Lauritzen (1990) presented a Bayesian analysis of acyclic directed graphical Markov models and this topic continues to attract research attention. Here we sketch the basic framework with a stylized version of a real epidemiological application.

In Norway, the Medical Birth Registry (MBR) gathers data nationwide on congenital malformations such as Down's syndrome. The primary purpose of the MBR is to track prevalences over time and identify abnormal trends. The data, however, are subject to a variety of errors, and epidemiologists have built statistical models to make inference about true prevalences. For Down's syndrome, such a model includes three dichotomous random variables: the reported Down's syndrome status, R, the true Down's syndrome status, S, and the maternal age, A, where age is dichotomized at 40, say.

Figure 1 displays a possibly reasonable model for these variables. This acyclic directed graph represents the assumption that the reported status and the maternal age are conditionally independent given the true status. The joint distribution of the three variables factors accordingly:

$$\Pr(A, S, R) = \Pr(A)\Pr(S \mid A)\Pr(R \mid S). \tag{1}$$

This factorization features a term for every vertex, the term being the conditional density of the vertex given its parents. In general, this factorization implies that

[1] This is somewhat of a misnomer since there is nothing Bayesian *per se* about Bayesian networks.

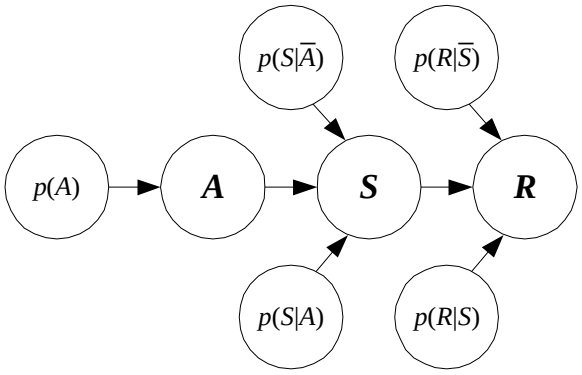


Figure 2: Down's Syndrome: An Acyclic Directed Bayesian Graphical Markov Model

vertices (more correctly, the random variables corresponding to vertices) are conditionally independent of their non-descendants given their parents (Lauritzen *et al.*, 1990).

The specification of the joint distribution of A, S, and R, in (1), requires five parameters:

$$\Pr(R \mid S), \Pr(R \mid \overline{S}), \Pr(S \mid A), \Pr(S \mid \overline{A}) \quad \text{and} \quad \Pr(A) \tag{2}$$

where $\overline{S}$, for example, denotes the absence of Down's syndrome. Once these probabilities are specified, the calculation of specific conditional probabilities such as $\Pr(R \mid A)$ can proceed via a series of local calculations without storing the full joint distribution (Dawid, 1992).

To facilitate Bayesian learning for the five parameters, Spiegelhalter and Lauritzen (1990) and Cooper and Herkovits (1992) make two key assumptions that greatly simplify subsequent analysis.

First, they assume that the parameters are independent *a priori*. Figure 2 embodies this assumption. For instance, $\text{pr}(S \mid A)$ in Figure 2 has no parents. Therefore, it is marginally independent of, for instance, $\text{pr}(A)$, since this is not a descendant of $\text{pr}(S \mid A)$.

Second, they assume that each of the probabilities has a beta distribution (or Dirichlet distribution for categorical variables with more than 2 levels).

Calculation of the posterior distributions is straightforward. Suppose we have the following prior distributions for three of the parameters: pr(A) $\sim$ beta(1,1), pr($S \mid A$) $\sim$ beta(5,1), and pr($R \mid \overline{S}$) $\sim$ beta(1,9), and we observe a single case, $d = (A, \overline{S}, R)$, that is, maternal age over 40 and incorrectly recorded by the MBR as Down's syndrome. Then, it is easy to show that pr(d) $= \frac{1}{1+1} \times \frac{1}{5+1} \times \frac{1}{1+9}$. Conditional on d, the posterior distributions of these parameters become: : pr(A) $\sim$ beta(2,1), pr($S \mid A$) $\sim$ beta(5,2), and pr($R \mid \overline{S}$) $\sim$ beta(2,9). The posterior distributions of the remaining two parameters, pr($S \mid \overline{A}$) and pr($R \mid S$), are unchanged. In this manner, we can sequentially calculate the likelihood for a collection of cases, D, conditional on the model of Figure 2. Heckerman *et al.* (1994) provide a closed-form expression for this likelihood.

We now discuss the actual Norwegian Down's syndrome example in more detail. Because of growing concerns about incomplete ascertainment, an additional 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. While it was expected that the MIA registration would be more accurate than the MBR, the MIA registration *is* subject to error. Table 1 presents data concerning Down's syndrome collected between 1985 and 1988 (Lie *et al.*, 1991,1994). The variables A and S continue to represent maternal age and true Down's syndrome status, respectively. R_1 represents case ascertainment through the national MBR registry and R_2 through the regional MIA registry.

Denoting by Q, the prevalence of Down's syndrome, and by Y, the observed data, the goal is to compute pr($Q \mid Y$). York *et al.* (1995) presented a Bayesian graphical Markov model analysis that addressed three substantive issues. First, there does exist prior knowledge about Down's syndrome rates, and the analysis sought to incorporate this . Second, different models exist for these data that provide a reasonable fit to the data, but lead to Down's syndrome prevalence estimates that are quite different. Third, the data are partially missing since R_2 is present for only a subset of the examples. Table 2 shows the results of analyses using various Bayesian graphical

Doubly Sampled Data

	A_1	A_2	A_3	A_4	A_5	A_6	Total
R_1, R_2	1	2	0	3	2	0	8
$R_1, \overline{R}_2$	5	2	3	0	2	1	13
$\overline{R}_1, R_2$	1	4	2	1	1	0	9
$\overline{R}_1, \overline{R}_2$	7478	10247	7058	2532	504	28	27847
Total	7485	10255	7063	2536	509	29	27877

Singly Sampled Data

	A_1	A_2	A_3	A_4	A_5	A_6	Total
R_1	32	55	58	62	23	3	233
$\overline{R}_1$	48957	70371	49115	16834	3348	165	188790
Total	48989	70426	49173	16896	3371	168	189023

Table 1: 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. A represents the maternal age in six categories: ≤ 24, 25—29, 30—34, 35—39, 40—44, and ≥ 45.

Markov models.

Each model used an informative prior distributions, based on historical data and expert knowledge for the various probabilities. Specifically, the parameters for the prior on pr(S), the prevalence of Down's syndrome, were based upon raw data from the MBR registry for 1979–1984 without any adjustment for misclassification. During that period, there were 0.97 observed cases per 1,000 births; we chose the beta$(0.0097, 9.9903)$ prior to have that rate as its expected value and so that our prior knowledge is equivalent to having observed 10 hypothetical births. The priors for the probability of a case being identified by the registries reflect our belief that the registries are more likely to find cases of the syndrome than not. For the national MBR registry, we placed beta(4,2) priors on pr($R_1 \mid S$) and pr($R_2 \mid S$). For the 6 parameters, pr($S \mid A_i$), we chose prior variances so that Var(p(S)) is the same

for all models, and prior expectations given by historical data for 1979–1984, which are $0.59, 0.59, 0.97, 3.04, 6.88$, and 18.50 cases per 1000 for age groups $A_1, \ldots, A_6$, as presented in Lie *et al.* (1991). York *et al.* (1995) describe the remaining prior distributions.

The analysis assumed 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. York *et al.* (1995) used a Markov chain Monte Carlo procedure to deal with the missing data (see Section 4).

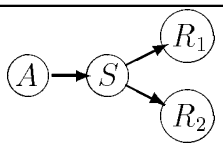
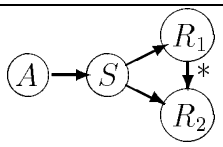
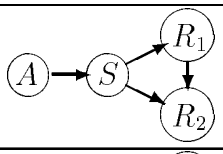
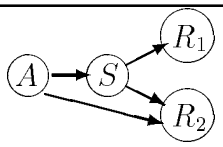
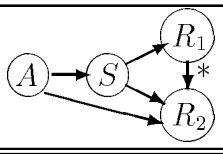
Model	Posterior Probability	$10^3 \times \Pr(S)$		
		Mode	Mean	Std Dev
	0.282	1.81	1.92	0.292
	0.385	1.49	1.51	0.129
	0.269	1.60	1.70	0.252
	0.030	1.71	1.78	0.226
	0.016	1.50	1.52	0.129

Table 2: Features of the posterior distribution for Down's syndrome prevalence. Prevalence is given as the rate per thousand. The "Posterior Probability" is a normalized marginal likelihood. Only models with posterior probability larger than 0.01 are listed.

Each model has an associated model score ("Posterior Probability" in Table 2) that reflects a tradeoff between how well the model fits the data and the complexity of the

model. It is clear that different plausible models can lead to quite different inferences. Section 3 discusses this issue in some detail.

2.3 Hierarchical Models and Exchangeability

For a data miner, the notion of "exchangeability" and the consequent models represent a powerful and underutilized tool. Here we introduce the basic idea by example.

Table 3 presents data concerning mortality rates in $k = 12$ hospitals performing cardiac surgery in babies. The analytic goal here is to rank order the hospitals according to their true mortality rates. Hospital A has the lowest observed mortality rate (0), but has performed the fewest surgeries (27). Should the analysis rank Hospital A as number one?

	Hospital											
	A	B	C	D	E	F	G	H	I	J	K	L
No. of Ops (n)	27	148	119	810	211	196	148	215	207	97	256	360
No. of Deaths (x)	0	18	8	46	5	13	9	31	14	8	29	24

Table 3: Hospital mortality data

Denote by r_i the true mortality rate for Hospital i. A Bayesian analysis must specify prior distributions for each of the r_i's. In the absence of any further information about the hospitals, these prior distributions must treat the r_i's symmetrically. In particular, in this situation, the Bayesian analysis would specify identical marginal prior distributions for each of the rates:

$$p(r_i) \equiv p(r_j) \forall i, j$$

as well as identical marginal joint prior distributions for all pairs of rates:

$$p(r_i, r_{i'}) \equiv p(r_j, r_{j'}) \forall i, i', j, j',$$

all triples, etc. Probabilistically we can represent this symmetry through *exchangeability*: the parameters $r_1, \ldots, r_k$ are exchangeable if their joint distribution $p(r_1, \ldots, r_k)$ is invariant to permutations of the indices $(1, \ldots, k)$.

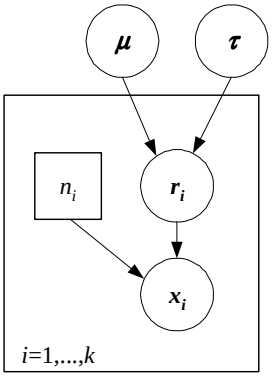


Figure 3: *Bayesian models from exchangeability.*

A remarkable result due to deFinetti says that in the limit as $k \to \infty$ any exchangeble sequence $r_1, \ldots, r_k$ can be written as:

$$p(r_1, \ldots, r_k) = \int \left[\prod_{i=1}^{k} p(r_i|\phi) \right] p(\phi) d\phi$$

for some $p(\phi)$. So, exchangeability suggests that we assume that the parameters $r_1, \ldots, r_k$ comprise independent and identically distributed draws from some "mother-distribution," denoted here by $p(\phi)$. In the cardiac surgery example, we could proceed as follows. Assume that within hospital i, the number of deaths x_i has a binomial distribution with parameters n_i (this is known) and r_i (this is unknown). Next we assume that the r_i's are independent and identically distributed where $\log(\frac{r_i}{1-r_i})$ is normally distributed with mean μ and precision τ. Finally we put fairly flat prior distributions on μ and τ. The Bayesian literature refers to such models *hierarchical models*. Figure 3 shows a graphical Markov model representation.

Figure 3 uses some established graphical conventions. The box around n_i indicates that this is a known quantity. The box around n_i, r_i, and x_i is called a *plate* indexed by $i = 1, \ldots, k$ and is a shorthand for repeating each n, r and x, k times. Every arrow entering the plate would also be repeated - for instance, there is an arrow from a to each of $r_1, \ldots, r_k$.

A Markov chain Monte Carlo algorithm (see Section 4) computes posterior distributions for r_i's. Table 4 shows the results.

	Hospital											
	A	B	C	D	E	F	G	H	I	J	K	L
No. of Ops (n)	27	148	119	810	211	196	148	215	207	97	256	360
Raw Rate (x/n)	0.00	12.16	6.72	5.68	2.37	6.63	6.08	14.42	6.76	8.25	11.33	6.67
Post. Mean	5.77	10.50	7.01	5.88	4.15	6.86	6.58	12.58	6.94	7.85	10.34	6.81
Post. S.D.	2.3	2.3	1.8	0.8	1.3	1.5	1.6	2.2	1.5	2.1	1.8	1.2
Raw Rank	1	11	7	3	2	5	4	12	8	9	10	6
Post. Rank	2	11	8	3	1	6	4	12	7	9	10	5

Table 4: Hospital mortality rate results (rates $\times$ 100)

The effect of the exchangeability assumption is to shrink the observed rates towards the overall average mortality rate (here 7.3%). For any given hospital, the amount of shrinkage depends on the number of surgeries performed. For example, Hospital D performed 810 surgeries. The mean of the posterior distribution for Hospital D's rate (5.76%) is close to the raw rate (5.68%). Hospital A on the other hand, performed only 27 surgeries. The posterior mean for Hospital A is 3.76% in contrast with raw rate of 0%. Indeed, ranking the hospitals according to their posterior means leads to the conclusion that Hospital E is the best hospital despite the fact that no deaths took place at Hospital A!

The phenomenon we are observing here is a sort of "borrowing strength." The hospitals with a paucity of data borrow inferential strength from the hospitals with abundance of data. The notion of exchangeability generalizes to partial exchangeability where observations form exchangeable subgroups (see, for example, Consonni and Veronese, 1995). Draper *et al.* (1993) provide an interesting discussion of the central role of exchangeability in data analysis (equally, in data mining).

2.4 Prior Distributions in Practice

We have already seen two examples involving the use of informative prior distributions. In practice, and especially in the context of large datasets, Bayesian analyses

often instead use prior distributions that reflect some degree of prior indifference or ignorance. The basic requirement for such a prior distribution is that it be flat in the region of the parameter space favored by the likelihood. Certainly it is always advisable to make sure that the conclusions are robust to the particular choice of prior. Carlin and Louis (2000, Section 6.1) provide a detailed discussion of different robustness analyses.

In an attempt to represent prior ignorance, some Bayesian analyses utilize so-called *improper priors*. These are priors that do not integrate to one but are uniform over, for instance, the real numbers. Hence these improper priors are not probability distributions at all. In fact, improper priors can actually lead to "improper posteriors," that is, posterior densities that do not integrate to one, an undesirable state of affairs. We recommend that data miners stick to proper prior distributions.

Concerning informative prior distributions, Chaloner (1996) surveys methodology for eliciting priors from human experts. Spiegelhalter *et al.* (1994) provide an exquisite example of pragmatic prior construction in the context of high-stakes decision making. Madigan *et al.* (1995) describe an example where informative priors improved predictive performance.

The *empirical Bayes* approach learns the prior distribution from the data. Carlin and Louis (2000) provide a thorough introduction. DuMouchel (1999) and DuMouchel and Pregibon (2001) describe applications of empirical Bayes to frequent itemset identification, both with methodology suitable for massive datasets.

3 Bayesian Model Selection and Model Averaging

The example of Section 2.2 raised a challenging and ubiquituous issue: often there are many competing "full probability models" that may all seem reasonable in the light of the prior knowledge and of the data, but lead to different predictions[2]. The Bayesian

[2]here we use the term "prediction" to mean inference for any unknown in the model, be it a future observable or not.

literature provides two general strategies to deal with this issue. *Model selection* chooses the single "best" model and bases subsequent predictions on that model. *Model averaging* combines the models and computes predictions as weighted averages. Both approaches have strengths and weaknesses. Model selection is computationally more straightforward and provides a single model to examine and critique. However, predictions that condition on a single model ignore model uncertainty and can be badly calibrated. Model averaging can be computationally demanding and inscrutable but usually provides better predictions.

3.1 Model Selection

Suppose a particular analysis leads to set of candidate models $\{M_i, i \in I\}$ and denote by $D = \{x_1, \ldots, x_n\}$ the observed data. The *posterior model probabilities*, $\{p(M_i|D), i \in I\}$ play a central role in Bayesian model selection. If we believe that *one* of the models $\{M_i, i \in I\}$ actually generated the data, then, if our goal is to choose the true model, the optimal decision is to choose the model with the highest posterior probability (Bernardo and Smith, 1994, p.389).

In practice, rather than trying to choose the "true" model, our goal is usually to make the best predictions possible. Furthermore, we don't usually believe that *any* of the models under consideration actually generated the data. Nonetheless, the posterior model probability is sometimes a reasonable choice as a model selection criterion (although not usually predictively optimal - see Barbieri and Berger, 2002). Note that:

$$p(M_i|D) \propto p(D|M_i) \times p(M_i)$$

so computing a posterior model probability involves two terms, the marginal likelihood and the prior model probability, as well as a normalization over the competing models. There exists a vast literature on marginal likelihood computation - we refer the interested reader to Section 6.3 of Carlin and Louis (2000) and to Han and Carlin (2001). Suffice it to say that the computation can sometimes prove challenging.

Bernardo and Smith (1984, Section 6.1.6) argue for selecting the model that maxi-

mizes a cross-validated predictive score. For the case of the logarithmic score, this selects the model that maximizes:

$$\frac{1}{k}\sum_{j=1}^{k}\log p(x_j|M_i,\mathbf{x}_{n-1}(j))$$

where $\mathbf{x}_{n-1}(j)$ represents the data with observation x_j removed, $\{x_1,\ldots,x_k\}$ represents a random sample from D, and the maximization is over $i \in I$.

For some applications, computational complexity may rule out cross-validatory predictive model selection procedures. Carlin and Louis (2000, Section 6.5.1) discuss alternatives. Spiegelhalter *et al.* (2002) propose a model selection criterion especially well suited to Bayesian hierarchical models and present a decision-theoretic justification. See also Geisser (1993) and Shibata (1983).

3.2 Model Averaging

If we believe that *one* of the models $\{M_i, i \in I\}$ actually generated the data, then, if our goal is to minimize the squared loss for future predictions, the optimal strategy is to *average* over all models (Bernardo and Smith, 1994, p.398). Specifically, if Q is the quantity of interest we wish to compute, Bayesian model averaging computes:

$$p(Q) = \sum_{i\in I} p(Q|D,M_i)p(M_i|D).$$

Bayesian model averaging is also optimal with regard to a logarithmic predictive score (Madigan and Raftery, 1994). As with model selection, we don't usually believe that one of the candidate models actually generated the data, but empirical evidence suggests that Bayesian model averaging usually provides better predictions than any single model (see, for example, Hoeting *et al.*, 1999), sometimes substantially better. Predictive distributions from Bayesian model averaging usually have bigger variances, more faithfully reflecting the real predictive uncertainty. Draper (1995) provides examples where failing to account for model uncertainty proved unfortunate.

Hoeting *et al.* (1999) discuss computing for Bayesian model averaging. Other references include Madigan and York (1995), Carlin and Chib (1995), and Carlin and Louis (2000, Section 6.4).

3.3 Model Assessment

Both model selectors and model averagers need tools to assess model fit. Bayesian residuals provide a useful starting point:

$$r_i = y_i - E(y_i|D), i = 1, \ldots, m$$

where $\{y_1, \ldots, y_m\}$ is a validation sample independent of the training data D, and E denotes expectation. Examination of these residuals can reveal failure in a distributional assumption or failure of some independence assumption, or whatever. Carlin and Louis (2000, Section 2.4) discuss a cross-validatory approach, wherein the fitted value for y_i is computed conditional on all the data except y_i, namely, $\mathbf{y}_{(i)} \equiv (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$, yielding:

$$r_i' = y_i - E(y_i|\mathbf{y}_{(i)}).$$

The *conditional predictive ordinate* (CPO), $p(y_i|\mathbf{y}_{(i)})$, is also quite useful. Individual data values having low CPO are poorly fit by the model. Gelfand and Dey (1984) point out the interesting fact that:

$$\frac{1}{p(y_i|\mathbf{y}_{(i)})} = \int \frac{1}{p(y_i|\mathbf{y}_{(i)}, \theta)} p(\theta|y) d\theta.$$

Hence a Monte Carlo estimate of the harmonic mean of $p(y_i|\mathbf{y}_{(i)}, \theta)$ estimates the CPO. This is usually straightforward to compute using, for example, BUGS (see 5.1 below). Harmonic means, however, can be numerically unstable.

Gelman *et al.* (1995, Chapter 6) favor a technique that involves drawing simulated values from the posterior predictive distribution of replicated data and comparing these samples to the observed data using some discrepancy measure. Different choices of discrepancy measure can shed light on different aspects of model fit. We refer the reader to Gelman *et al.* for further details.

4 Bayesian Computation

Outputs of Bayesian data analyses often come in the form of expectations such as the marginal means, variances, and covariances of the quantity of interest. We compute the expected value of the quantity of interest, $h(\theta)$, using

$$E(h(\theta)|x_1, \ldots, x_N) = \int h(\theta) f(\theta|x_1, \ldots, x_N) d\theta \tag{3}$$

where $f(\theta|\mathbf{x})$, is the posterior distribution of the parameters given the observed data. Computation of these expectations requires calculating integrals that, for all but the simplest examples, are difficult to compute in closed form. Monte Carlo integration methods sample from the posterior, $f(\theta|\mathbf{x})$, and appeal to the law of large numbers to estimate the integrals,

$$\lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} h(\theta_i) = \int h(\theta) f(\theta|x_1, \ldots, x_N) d\theta \tag{4}$$

where the θ_i compose a sample from $f(\theta|\mathbf{x})$.

The ability to compute these expectations efficiently is equivalent to being able to sample efficiently from $f(\theta|\mathbf{x})$. Sampling schemes are often difficult enough without the burden of large datasets. The additional complexity of massive datasets usually causes each iteration of the Monte Carlo sampler to be slower. When the number of iterations already needs to be large, efficient procedures within each iteration are essential to timely delivery of results. So-called "variational approximations" can be useful for massive datasets beyond the reach of Monte Carlo methods and we discuss those briefly at the end of this Section.

4.1 Importance sampling

Importance sampling is a general Monte Carlo method for computing integrals. As previously mentioned, Monte Carlo methods approximate integrals of the form (3). The approximation in (4) depends on the ability to sample from $f(\theta|\mathbf{x})$. When a sampling mechanism is not readily available for the "target distribution," $f(\theta|\mathbf{x})$, but

one is available for another "sampling distribution," $g(\theta)$, we can use importance sampling. Note that for (3) we can write

$$\int h(\theta) f(\theta|x_1, \ldots, x_N) d\theta = \int h(\theta) \frac{f(\theta|\mathbf{x})}{g(\theta)} g(\theta) d\theta \quad (5)$$

$$= \lim_{M \to \infty} \sum_{i=1}^{M} w_i h(\theta_i) \quad (6)$$

where θ_i is a draw from $g(\theta)$ and $w_i = f(\theta_i|\mathbf{x})/g(\theta_i)$. Note that the expected value of w_i under $g(\theta)$ is 1. Therefore, if we are able to compute the importance sampling weights, w_i, only up to a constant of proportionality, we can normalize the weights to compute the integral.

$$\int h(\theta) f(\theta|x_1, \ldots, x_N) d\theta = \lim_{M \to \infty} \frac{\sum_{i=1}^{M} w_i h(\theta_i)}{\sum_{i=1}^{M} w_i} \quad (7)$$

Naturally, in order for the sampling distribution to be useful, drawing from $g(\theta)$ must be easy. We also want our sampling distribution to be such that the limit converges quickly to the value of the integral. If the tails of $g(\theta)$ decay faster than $f(\theta|\mathbf{x})$ the weights will be numerically unstable. If the tails of $g(\theta)$ decay much more slowly than $f(\theta|\mathbf{x})$ we will frequently sample from regions where the weight will be close to zero, wasting computation time. Second to sampling directly from $f(\theta|\mathbf{x})$, we would like a sampling distribution slightly fatter than $f(\theta|\mathbf{x})$.

In Ridgeway and Madigan (2002) we show that when we set the sampling density to be $f(\theta|x_1, \ldots, x_n)$, where $n \ll N$ so that we condition on a manageable subset of the entire dataset, the importance weights for each sampled θ_i require only one sequential scan of the remaining observations. Before beginning that discussion, the next section introduces the most popular computational method for Bayesian analysis of complex models.

4.2 Markov chain Monte Carlo (MCMC)

Importance sampling is a useful tool, but for complex models, crafting a reasonable sampling distribution can be extremely difficult. Gilks *et al.* (1996) provides a more

detailed introduction to MCMC along with a variety of interesting examples and applications.

As with importance sampling, the goal is to generate a set of draws from the posterior distribution $f(\theta|\mathbf{x})$. Rather than create independent draws and reweight, MCMC methods build a Markov chain, a sample of *dependent* draws, $\theta_1, \ldots, \theta_M$, that have stationary distribution $f(\theta|\mathbf{x})$. It turns out that it is often easy to create such a Markov chain with a few basic strategies. However, there is still a bit of art involved in creating an efficient chain and assessing the chain's convergence.

Figure 4 shows the Metropolis-Hastings algorithm (Hastings, 1970), a very general MCMC algorithm. Assume that we have a single draw θ_1 from $f(\theta|\mathbf{x})$ and a proposal density for a new draw, $q(\theta|\theta_1)$. If we follow step 2 of the MCMC algorithm then the distribution of θ_2 will also be $f(\theta|\mathbf{x})$. This is one of the key properties of the algorithm. Iterating this algorithm we will obtain a sequence $\theta_1, \ldots, \theta_M$ that has $f(\theta|\mathbf{x})$ as its stationary distribution.

MCMC methods have two main advantage that make them so useful for Bayesian analysis. First, we can choose q's from which it is easy to simulate. Special choices for q, which can depend on the data, simplify the algorithm. If q is symmetric, for example a Gaussian centered on θ_{i-1}, then the entire proposal distributions cancel out in (8). If we choose a q that proposes values that are very close to θ_{i-1} then it will almost always accept the proposal but the chain will move very slowly and take a long time to converge to the stationary distribution. If q proposes new draws that are far from θ_{i-1} and outside the region with most of the posterior mass, the proposals will almost always be rejected and again the chain will converge slowly. With a little tuning the proposal distribution can usually be adjusted so that proposals are not rejected or accepted too frequently. Essentially the only constraint on the choice of q is that it results in an irreducible and aperiodic chain. The second advantage is that there is no need to compute the normalization constant of $f(\theta|\mathbf{x})$ since it cancels out in (8).

The Gibbs sampler (Geman and Geman, 1984) is a special case of the Metropolis-Hastings algorithm and is especially popular. If θ is a multidimensional parameter,

the Gibbs sampler sequentially updates each of the components of θ from the full conditional distribution of that component given fixed values of all the other components and the data. For many models used in common practice, even the ones that yield a complex posterior distribution, sampling from the posterior's full conditionals is often a relatively simple task. Conveniently, the acceptance probability (8) always equals one.

4.3 An example

Consider again the hospital mortality example of Section 2.3. Recall that the unknowns of are the hospital mortality rates, $r_i, i = 1, \ldots, 12$, For simplicity here we will assume that the r_i's come from a beta distribution with parameters a and b and the analyst has endowed a and b with some prior distribution, say $p(a, b)$. Gibbs sampling here requires that we sample from the conditional distribution of each parameter, given all the other parameters and given the observed data. The trick here is to first write down the joint density of all the random quantities in the model:

$$p(r_1, \ldots, r_{12}, x_1, \ldots, x_{12}, a, b) = \prod_{i=1}^{12} \{p(x_i|r_i, n_i)p(r_i|a, b)\}p(a, b).$$

Since:

$$p(r_i|r_1, \ldots, r_{i-1}, r_{i+1}, \ldots, r_{12}, x_1, \ldots, x_{12}, a, b) \propto p(r_1, \ldots, r_{12}, x_1, \ldots, x_{12}, a, b),$$

we have that:

$$p(r_i|r_1, \ldots, r_{i-1}, r_{i+1}, \ldots, r_{12}, x_1, \ldots, x_{12}, a, b) \propto p(x_i|r_i, n_i)p(r_i|a, b)$$

(that is, just pick off the terms in the joint density that involve r_i). Since x_i is a binomial random variable and r_i has a beta prior distribution, we have:

$$p(r_i|r_1, \ldots, r_{i-1}, r_{i+1}, \ldots, r_{12}, x_1, \ldots, x_{12}) \sim \text{Beta}(a + x_i, b + n_i - x_i).$$

Thus the Gibbs sampler here simply cycles through the twelve rates in any order, sampling each time from a beta distribution. The Gibbs sampler will also sample from a and b, the specific distributions depending on the choice of $p(a, b)$.

Missing data fits into this scheme in a straightforward fashion. Suppose x_5 is missing. Then the Gibbs sampler expands to include a draw from the conditional distribution of x_5 given everything else. The veracity of this approach does depend on somewhat subtle assumptions about the missing data mechanism. See Gelman *et al.* (1995, Chapter 17) for a detailed discussion.

4.4 Application to Massive Data

MCMC as specified, however, is computationally infeasible for massive datasets. Except for the most trivial examples, computing the acceptance probability (8) requires a complete scan of the dataset. Although the Gibbs sampler avoids the acceptance probability calculation, precalculations for simulating from the full conditionals of $f(\theta|\mathbf{x})$ require a full scan of the dataset, sometimes a full scan for each component! Since MCMC algorithms produce dependent draws from the posterior, M usually has to be very large to reduce the amount of Monte Carlo variation in the posterior estimates. While MCMC makes fully Bayesian analysis practical it seems dead on arrival for massive dataset applications.

Although this section has not given great detail about the MCMC methods, the important ideas for the purpose of this Chapter are that

1. MCMC methods make Bayesian analysis practical,
2. MCMC often requires an enormous number of laps through the dataset, and
3. given a θ drawn from $f(\theta|\mathbf{x})$ we can use MCMC to draw another value, θ', from the same distribution.

4.5 Importance sampling for analysis of massive datasets

So far we have two tools, importance sampling and MCMC, to draw samples from an arbitrary posterior distribution. Ridgeway and Madigan (2002) present a partic-

1. Initialize the parameter θ_1

2. For i in $2, \ldots, M$ do
   Step (a) and/or (b) requires a scan of the dataset

   (a) Draw a proposal θ' from $q(\theta|\theta_{i-1})$,

   (b) Compute the acceptance probability

$$\alpha(\theta', \theta_{i-1}) = \min\left(1, \frac{f(\theta'|\mathbf{x})q(\theta_{i-1}|\theta')}{f(\theta_{i-1}|\mathbf{x})q(\theta'|\theta_{i-1})}\right) \tag{8}$$

   (c) With probability $\alpha(\theta', \theta_{i-1})$ set $\theta_i = \theta'$.
   Otherwise set $\theta_i = \theta_{i-1}$

Figure 4: The Metropolis-Hastings algorithm

ular form of importance sampling that helps perform Bayesian analysis for massive datasets.

Ideally we would like to sample efficiently and take advantage of all the information available in the dataset. A factorization of the integrand shows that this is possible when the observations, x_i, are exchangeable. Let D_1 and D_2 be a partition of the dataset so that every observation is in either D_1 or D_2.

As noted for (3) we would like to sample from the posterior conditioned on all of the data, $f(\theta|D_1, D_2)$. Since sampling from $f(\theta|D_1, D_2)$ is difficult due to the size of the dataset, we consider setting $g(\theta) = f(\theta|D_1)$ for use as our sampling distribution and using importance sampling to adjust the draws. If θ_i, $i = 1, \ldots, M$, are draws from $f(\theta|D_1)$ then we can estimate the posterior expectation (3) as

$$\hat{E}(h(\theta)|D_1, D_2) = \frac{\sum_{i=1}^{M} w_i h(\theta_i)}{\sum_{i=1}^{M} w_i} \tag{9}$$

where the w_i's are the importance sampling weights

$$w_i = \frac{f(\theta_i|D_1, D_2)}{f(\theta_i|D_1)}. \tag{10}$$

Although these weights still involve $f(\theta_i|D_1, D_2)$, they greatly simplify.

$$\begin{aligned} w_i &= \frac{f(D_1, D_2|\theta_i)f(\theta_i)}{f(D_1, D_2)} \frac{f(D_1)}{f(D_1|\theta_i)f(\theta_i)} && (11) \\ &= \frac{f(D_1|\theta_i)f(D_2|\theta_i)f(D_1)}{f(D_1|\theta_i)f(D_1, D_2)} && (12) \\ &= \frac{f(D_2|\theta_i)}{f(D_2|D_1)} \\ &\propto f(D_2|\theta_i) = \prod_{x_j \in D_2} f(x_j|\theta_i) && (13) \end{aligned}$$

Line (11) follows from applying Bayes' theorem to the numerator and denominator. Equation (12) follows from (11) since the observations are assumed to be exchangeable, the observations in the dataset partition D_1 are independent from those in D_2 given θ. Conveniently, (13) is just the likelihood of the observations in D_2 evaluated at the sampled value of θ. Figure 5 summarizes this result as an algorithm. The algorithm maintains the weights on the log scale for numerical stability.

So rather than sample from the posterior conditioned on all of the data, D_1 and D_2, which slows the sampling procedure, we need only sample from the posterior conditioned on D_1. The remaining data, D_2, simply adjusts the sampled parameter values by reweighting. The for loops in step 5 of figure 5 are interchangeable. The trick here is to have the inner loop scan through the draws so that the outer loop only needs to scan D_2 once to update the weights. Although the same computations take place, in practice physically scanning a massive dataset is far more expensive than scanning a parameter list. However, massive models as well as massive datasets exist so that in these cases scanning the dataset may be cheaper than scanning the sampled parameter vectors. We will continue to assume that scanning the dataset is the main impediment to the data analysis.

We certainly can sample from $f(\theta|D_1)$ more efficiently than from $f(\theta|D_1, D_2)$ since simulating from $f(\theta|D_1)$ will require a scan of a much smaller portion of the dataset. For reasons discussed later, the algorithm works best when D_1 is as large as possible.

1. Load as much data into memory as possible to form D_1, taking into account space requirements for the Monte Carlo algorithm
2. Draw M times from $f(\theta|D_1)$ via Monte Carlo or Markov chain Monte Carlo
3. Purge the memory of D_1
4. Create a vector of length M to store the logarithm of the weights and initialize them to 0
5. Iterate through the remaining observations. For each observation, x_j, update the log-weights on all of the draws from $f(\theta|D_1)$

   for x_j in the partition D_2 do
   {
   for i in $1, \ldots, M$ do
   {
   $\log w_i \leftarrow \log w_i + \log f(x_j|\theta_i)$
   }
   }
6. Rescale to compute the weights

$$w_i \leftarrow \exp\left(\log w_i - \max(\log w_i)\right) \tag{14}$$

Figure 5: Importance sampling for massive datasets

We also assume that, for a given value of θ, the likelihood is readily computable up to a constant, which is almost always the case. When some data are missing, the processing of an observation in D_2 will require integrating out the missing information. Since the algorithm handles each observation case by case, computing the observed likelihood as an importance weight will be much more efficient than if it was embedded and repeatedly computed in a Metropolis-Hastings rejection probability computation. Placing observations with missing values in D_2 greatly reduces the number of times this integration step needs to occur, easing likelihood computations.

Ridgeway and Madigan (2002) describe some of the limitations of the algorithm shown in Figure 5 and describe a more elaborate algorithm based in recent work on sequential Monte Carlo algorithms. We refer the interested reader to Ridgeway and Madigan (2002) and to Chopin (2002) for details.

4.6 Variational Methods

Variational approximations for Bayesian calculations have emerged in recent years as a viable alternative to MCMC. The variational approach scales well and can tackle problems beyond the reach of MCMC (see, for example, Blei *et al.*, 2003). Here, following Jaakola and Jordan (2000), we sketch the basic idea in the specific context of logistic regression. Recall that a logistic regression takes the form:

$$p(Y = 1|\mathbf{X}, \theta) = g(\theta^T \mathbf{X})$$

where $g(\theta^T X) = (1 + e^{-x})^{-1}$. Given training data indexed by $i = 1, \ldots, n$, the posterior distributution of θ is:

$$p(\theta|\{Y_i, \mathbf{X}_i\}_{i=1}^n) \propto p(\theta) \prod_{i=1}^{n} \Big(Y_i \times g(\theta^T \mathbf{X}_i) + (1 - Y_i) \times (1 - g(\theta^T \mathbf{X}_i)) \Big)$$

where $p(\theta)$ is some prior distribution for θ.

The first step is to symmetrize the logistic function:

$$\log g(x) = -\log(1 + e^{-x}) = \frac{x}{2} - \log(e^{x/2} + e^{-x/2})$$

and note that $f(x) = -\log(e^{x/2} + e^{-x/2})$ is a convex function in the variable x^2. Thus it is possible to bound f below by a function that is linear in x^2. Specifically:

$$\begin{aligned} f(x) &\geq f(\xi) + \frac{\partial f(\xi)}{\partial(\xi^2)}(x^2 - \xi^2) && (15) \\ &= -\xi/2 + \log g(\xi) + \frac{1}{4\xi} \tanh(\xi/2)(x^2 - \xi^2). && (16) \end{aligned}$$

This bound is exact when $\xi^2 = x^2$. The essential idea of the variational approximation is to use this lower bound in place of $\log g()$ in the calculation of the posterior

distribution of θ. Jaakola and Jordan (2000) use an EM algorithm to find the value for ξ that makes the bound as tight as possible across all the training data. Since the bound is quadratic in x, choosing, for instance, a Gaussian prior for θ will yield a closed form expression for the approximate posterior distribution. In a sense, the variational approach here replaces an integration problem by a simpler optimization problem.

5 Bayesian modeling

We begin with a discussion of BUGS, a singular general purpose software tool for Bayesian data analysis. Then we briefly discuss specific Bayesian models and provide pointers to the literature.

5.1 BUGS and Models of realistic complexity via MCMC

BUGS is a terrific tool for Bayesian data mining. The UK Medical Research Council at Cambridge developed BUGS over the last decade. The program is available free-of-charge from:

`http://www.mrc-bsu.cam.ac.uk/bugs/`.

There are versions for Unix, DOS, and Windows (WinBUGS). The BUGS manual (Spiegelhalter *et al.*, 1999) describes BUGS:

> "BUGS is a computer program that carries out Bayesian inference on statistical problems using Gibbs sampling.
>
> BUGS assumes a Bayesian or full probability model, in which all quantities are treated as random variables. The model consists of a defined joint distribution over all unobserved (parameters and missing data) and observed quantities (the data); we then need to condition on the data

in order to obtain a posterior distribution over the parameters and unobserved data. Marginalising over this posterior distribution in order to obtain inferences on the main quantities of interest is carried out using a Monte Carlo approach to numerical integration (Gibbs sampling).

There is a small set of BUGS commands to control a session in which a (possibly very complex) statistical model expressed using the BUGS language is analysed. A compiler processes the model and available data into an internal data structure suitable for efficient computation, and a sampler operates on this structure to generate appropriate values of the unknown quantities.

BUGS is intended for complex models in which there may be many unknown quantities but for which substantial conditional independence assumptions are appropriate. Particular structures include generalised linear models with hierarchical or crossed random effects, latent variable or frailty models, measurement errors in responses and covariates, informative censoring, constrained estimation, and missing data."

For example, Figure 6 shows the BUGS code for the cardiac surgery example of Section 2.3. Alternatively and equivalently, WinBUGS can represent the model graphically. Figure 7 shows a screendump. Here we selected the node $r[i]$ and the two lines at the top of the figure show the node's specification. We see that it is a *stochastic node* (represented by an oval) with a binomial distribution conditional on its parents in the graph which are the binomial parameters $p[i]$ and $n[i]$. The two other node types are *constant nodes* (represented by rectangles) like a, b, and $n[i]$, or *logical nodes* (represented by ovals with dashed incoming edges – see Figure 9) which are functions of other nodes.

The primary BUGS output comprises summary statistics and density estimates for all the unknowns. BUGS can also provide the MCMC iterates for input to MCMC diagnostic programs like CODA (CODA is distributed with BUGS). Figure 8 shows the density estimates for the first four hospitals in the cardiac surgery example.

```
model;
const
    N = 12,                    # number of hospitals
    a = 2, b = 24;             # hyperparameters
var
    r[N],                      # number of deaths
    n[N],                      # total number of operations
    p[N];                      # true probability of death

data r,n in 'surgical.dat';

{
    for( i in 1 : N ) {
        r[i] ~ dbin(p[i],n[i])
        p[i] ~ dbeta(a,b)
    }
}
```

Figure 6: BUGS code for the cardiac surgery example

The combination of the graphical Markov model framework, the Bayesian approach, and MCMC means that BUGS can build models of considerable complexity, when such complexity is warranted. To demonstrate this, we present here an application of moderate complexity that would challenge the capabilities of any commercial data analysis or data mining software. This example is from Draper and Madigan (1997).

In the US, the anti-abortion campaign of the late 1980s and early 1990s generated much publicity and occasionally became violent. Sociologists, epidemiologists, and medical researchers have begun to study the effect that this campaign has had on access to reproductive services and on pregnancy terminations. Interest mainly lies in modeling the incidence rates of pregnancy terminations and their changes over time

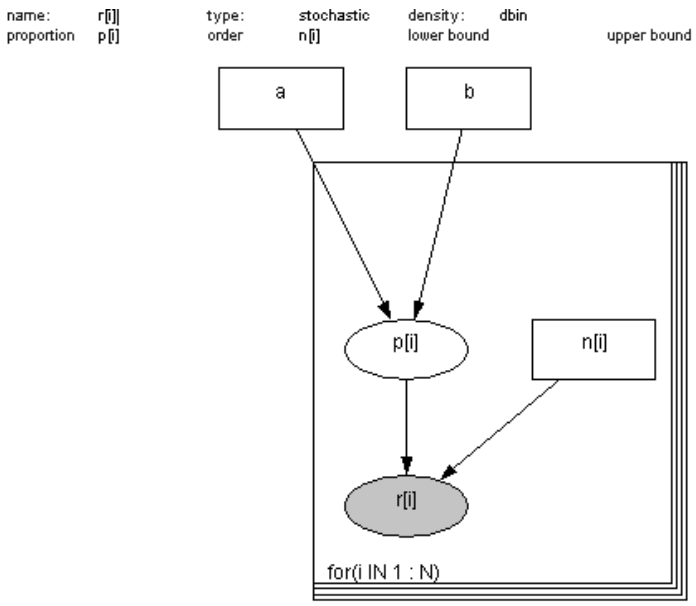


Figure 7: BUGS Graphical Markov Model for the cardiac surgery model.

at the county level, with particular attention focusing on possible changes in response to the anti-abortion campaign. Available data include the observed number y_{ijk} of reported terminations in age group i, county j, and year k (across 5 age groups, 38 US counties, and the years 1983–94), together with the appropriate population denominators n_{ijk} and a selection of county-level predictor variables x_j including, e.g., the estimated number of clinics providing reproductive services in 1983 and in 1993. A natural model for count data like the y_{ijk} involves regarding the counts as realizations of Poisson random variables,

$$y_{ijk} \sim Poisson(\mu_{ijk}), \tag{17}$$

where the mean structure μ_{ijk} is related to age group, district, and year. Preliminary exploratory data analysis and examination of several simple models suggests the structure

$$\log \mu_{ijk} = \log n_{ijk} + \alpha_i^a + \alpha_j^c + (\beta_i^a + \beta_j^c)t_k + (\gamma_i^a + \gamma_j^c)z_k, \tag{18}$$

where z_k is an indicator variable for after–1990 versus before and, e.g., α_i^a, β_i^a, and γ_i^a are the age effect on the intercept, the change per year, and the gradient after 1990. However, in the light of the data, several deficiencies in this model present themselves and prompt the following elaborations:

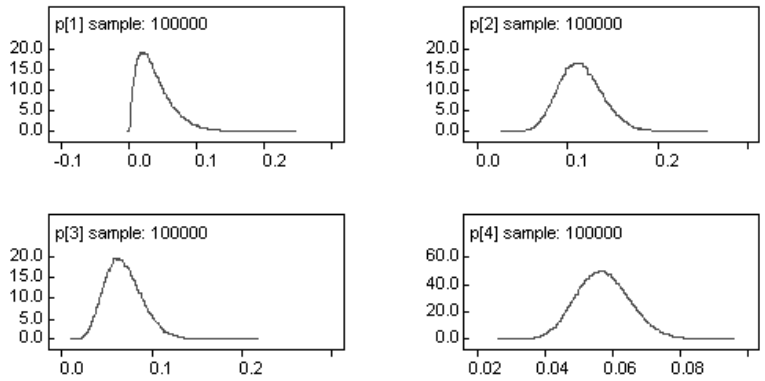


Figure 8: BUGS output for Hospitals A-D in the cardiac surgery model.

- *Over-dispersion*: The theoretical mean and variance of the Poisson distribution are equal, but with data such as these, sample variances often greatly exceed sample means. Thus the model typically needs to be generalized to allow for this extra variability, which can be thought of as arising from the omission of additional unobserved predictor variables. This generalization may be accomplished by adding a Gaussian error term to equation (18).

- *Hierarchical modeling*: The standard Bayesian modeling of the county-level coefficients α_j^c, β_j^c, and γ_j^c would involve specifying prior information—in the form of probability distributions—for each of them (i.e., 114 separate distributions). However, considering the counties as *exchangeable* pools the knowledge across them. This is accomplished mathematically by regarding, e.g., the 38 α_j^c's as having arisen from a single (e.g., Gaussian) distribution. That single distribution also may have a set of unknown hyperparameters that we could choose or estimate.

- *Relationships among the coefficients*: Prior experience suggests that it is unrealistic to consider the county-level coefficients α_j^c, β_j^c, and γ_j^c as independent, so the three distributions just described (one for each of α, β, and γ) need to be treated as correlated.

- *Underreporting*: Much is known about abortion underreporting rates, which is a form of missing data. It is necessary to incorporate this information into the model in a way that fully fleshes out the resulting uncertainty.

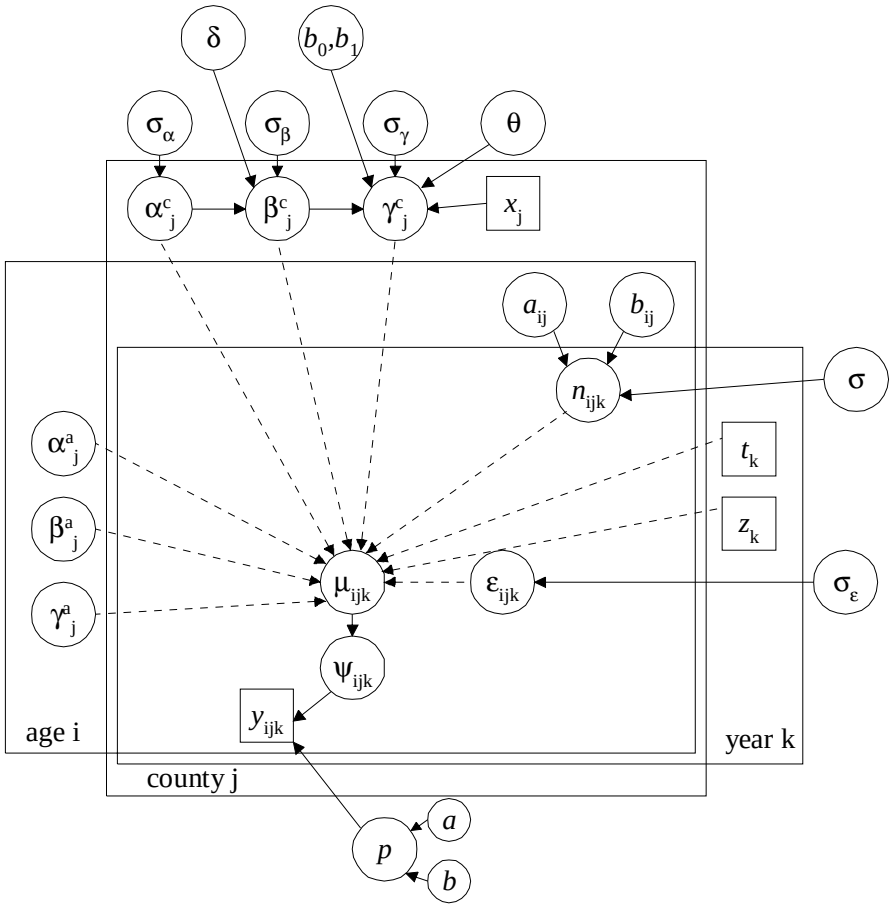


Figure 9: *Bayesian model for the reproductive services example.*

Figure 9 presents a summary of the resulting model. MCMC computation makes calculations for this Bayesian model straightforward. Moreover, complicated functions of the unknowns, such as the true relative rankings in abortion rates among the counties, are available trivially, together with measures of uncertainty (difficult to achieve correctly with non-Bayesian analyses) that fully reflect what is known and what is not.

See Spiegelhalter (1988) for another example along these lines.

The text by Congdon (2001) discusses Bayesian data analysis with BUGS and WinBUGS and provides sample code. The BUGS documentation itself features two volumes of examples covering a broad range of applications.

Note that BUGS does not provide any facilities for learning model structure. That is, the user specifies the graphical Markov model and then BUGS carries out the corresponding MCMC calculations. A BUGS module for learning model structure would be a useful addition.

5.2 Bayesian Predictive modeling

Thus far, we have dealt mostly with full probability modeling. Many applications, however, single out one quantity, y, and characterize y as a function of another quantity or vector of quantities x. As such, these applications want the conditional distribution of y given x, parameterized as $p(y|\theta, x)$ say, under a model in which the n observations $(x, y)_i$ are exchangeable (Gelman *et al.*, 1995, p.233). The quantity y calls itself the *response* or *outcome variable*. The variables $x = \{x_1, \ldots, x_k\}$ are the *explanatory variables* or *predictors*.

Gelman *et al.* (1995, p.235) note that a full probability model would also specify a distribution for x, say $p(x|\psi)$ for some parameter vector ψ, leading to the full probability model $p(x, y|\theta, \psi)$. However, if θ and ψ are independent in their prior distribution, that is, $p(\theta, \psi) = p(\theta)p(\psi)$, then the posterior distribution factors as:

$$p(\theta, \psi|x, y) = p(\psi|x)p(\theta|x, y)$$

and we can analyze the second factor by itself with no loss of information:

$$p(\theta|x, y) \propto p(\theta)p(y|x, \theta).$$

The practical advantage of this is that it is often much easier to devise a good conditional probability model for a single quantity than to come up with a full probability model for all quantities.

Many predictive applications use the *normal linear model* in which the distribution of y given x is a normal whose mean is a linear function of x:

$$E(y_i|\beta, x) = \beta_1 x_{i1} + \ldots + \beta_k x_{ik},$$

for $i = 1, \ldots, n$. Gelman *et al.* (1995, Chapter 8) discuss the Bayesian analysis of the normal linear model with a variety of variance structures. See also, Raftery *et al.* (1997).

Gelman *et al.* (1995) also discuss Bayesian analysis of hierarchical linear models and generalized linear models. Carlin and Louis (2000) include spatial and spatio-temporal models, nonlinear models, and longitudinal data models. West and Harrison (1997) is the classical reference on the Bayesian analysis of time series data. The companion text Pole *et al.* (1994) covers applications and software.

Survival analysis concerns the situation where y is a lifetime and, invariably, subject to various forms of censoring. Ibrahim *et al.* (2001) provide a thorough introduction to survival analysis from the Bayesian perspective. See also, Raftery *et al.* (1996) and Volinsky *et al.* (1997).

Neal (1996) provides an excellent description of the Bayesian analysis of neural networks. See also the extensive material on David MacKay's web site:

`http://www.inference.phy.cam.ac.uk/mackay/`.

Tree-based predictive models partition the predictor space into a set of rectangles, and then fit a simple model (like a constant) in each one. Tree models are conceptually simple, handle both discrete and continuous responses and predictors, and elegantly deal with missing values. However, tree models suffer from an instability problem, namely that a small change in the data can result in the tree-learning algorithm selecting a very different tree. Averaging procedures such as bagging or boosting effectively address the instability issue and can yield outstanding predictive accuracy. Hastie *et al.* (2001, Section 9.2) provide an introduction and references. Breiman and Friedman (1984) providing a fascinating early account of scaling tree-learning algorithms to massive data.

Chipman *et al.* (1998) and Denison *et al.* (1998) describe Bayesian tree models and associated MCMC learning algorithms. Chipman *et al.* (2001) report impressive predictive performance for their Bayesian tree algorithm. Bayesian model averaging

may address the instability problem associated with tree models, but this has yet to be fully explored. Denison *et al.* (2002) describe a related class of Bayesian partition models. The "partitions" in their case refer to partitions of continuous predictors. The method seems to scale reasonably well and provided good predictive performance on a banking application.

Readers may be interested in the (non-Bayesian) Microsoft Research "WinMine Toolkit." This software learns graphical models from data, using a decision tree to model the conditional distribution at each node. See:

`http://research.microsoft.com/~dmax/WinMine/Tooldoc.htm`.

We note that Bayesian predictive modelers can short-circuit some of the computational complexity of the fully Bayesian approach by making so-called "Bayesian plug-in" predictions. The essential idea to make predictions at a single parameter value, typically the posterior mode, instead of integrating over the posterior distribution. In practice, predictive accuracy is usually robust to this approximation and even outperforms the fully-Bayesian approach on occasion.

5.3 Bayesian Descriptive modeling

Earlier Sections presented several examples of Bayesian graphical Markov models and many commercial and non-commercial tools (such as BUGS) exist for learning such models from data. Kevin Murphy maintains a list at:

`http://www.cs.berkeley.edu/~murphyk/Bayes/bnsoft.html`.

Of course, graphical Markov models can be used predictively simply by considering conditional distributions. Naive Bayes models represent an important sub-class of graphical models that scale well and often yield surprisingly good predictive performance (see Hand and Yu, 2001, or Lewis, 1998). The classical Naive Bayes model (for example, Duda and Hart, 1973) imposes a conditional independence constraint,

namely that the predictor variables, say, $x_1, \ldots, x_k$, are conditionally independent given the response variable y. Figure 10 shows a graphical Markov model representation of the naive Bayes model.

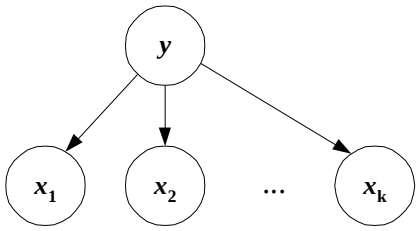


Figure 10: *The naive Bayes model.*

Outlier detection is an important data mining task and many authors have presented Bayesian approaches. A key advantage of the Bayesian approach is that it directly computes the probability that an observation is an outlier. See Hoeting *et al.* (1996), Justel and Pena (1999), Bayarri and Morales (2000), and the references therein.

Cluster analysis has developed mainly as a set of ad hoc methods. More recently, many data analysts have found that basing cluster analysis on a full probability model yields significant advantages. In particular, the probabilistic approach enables clustering of non-standard objects (such as web-page visits over time or gene expression data), can detect clusters within clusters or overlapping clusters, and can make formal inference about the number of clusters. Reversible jump MCMC (Green, 1995), an MCMC algorithm than incorporates jumps between spaces of varying dimension, provides a flexible framework for Bayesian analysis of model-based clustering. Richardson and Green (1997) is key a reference. See also Fraley and Raftery (1998) and the MCLUST software available from Raftery's website:

`http://www.stat.washington.edu/raftery/Research/Mclust/mclust.html`,

the AutoClass system (Cheeseman and Stutz, 1996), and the SNOB system (Wallace and Dowe, 2000). Cadez and Smyth (1999) and Cadez *et al.* (2000) present an EM algorithm for model-based clustering and describe several applications.

6 Available Software

Section 5.1 described the BUGS software. Carlin and Louis (2000, Appendix C) present a comprehensive guide to Bayesian software. Berger (2000) includes a list of Bayesian software websites as well as an overview of Bayesian philosophy and approaches:

`http://www.stat.duke.edu/~berger/papers/99-30.ps`

7 Discussion and Future Directions

Markov chain Monte Carlo has opened the floodgates to serious Bayesian data analyses. For example, the *Sixth Workshop on Case Studies of Bayesian Statistics* took place in 2001 and presented a broad array of real-world applications:

`http://lib.stat.cmu.edu/bayesworkshop/2001/Bayes01.html`.

While data mining applications often feature massive datasets, predictive modelers often deal with a paucity of *labeled* data. Bayesian analyses, however, can readily absorb external information sources and usefully bootstrap the learning process. For example, much of the empirical work in text categorization research uses massive, publicly-available test collections of labeled documents. The Reuters Corpus of news articles, for instance, contains over 800,000 labeled documents. Incorporation of prior information in such a context yields modest returns. Yet, real-world applications rarely feature large collections of labeled documents (Lewis, 2002) and prior knowledge can play a critical role. We suspect the the same is true of many real-world prediction problems.

Breiman (2001) argues that attempting to model mother nature's data generating mechanism is largely a futile exercise. He makes a distinction between the "data modeling culture" (this includes the full probability modelers and the Bayesians) and

the "algorithmic modeling culture." The algorithmic culture regards the data generating mechanism as complex and unknown, and focuses on algorithms, often very complicated algorithms, that yield good predictive performance. Breiman's distinction is perhaps not as sharp as he contends. For one thing, many of the algorithmic methods such as neural networks and support vector machines yield natural Bayesian analogues. See, for example, the *Gaussian Process* models of Williams (1998), or the *Bayes Point Machines* of Herbrich *et al.*, 2001). In fact, Herbrich *et al.* present Bayes point machines as an algorithmic approximation to Bayesian inference for kernel-based predictive models.

Combining multiple complex hierarchical models such as that in Figure 9, with model structures and priors tuned to optimize predictive performance or some other utility function, combines aspects of the data modeling culture and the algorithmic culture. For *multilevel data* such combinations seem especially expedient. For example, in a medical context, we may have information about individual patients, information about the hospitals in which they stayed, and information about the geographic locales of the hospitals. Assuming exchangeability or partial exchangeability within each level can sharpen inference and provide better predictions. Between the levels, the Bayesian literature usually features linear or mildly non-linear parametric models. Yet, many of the popular algorithmic or kernel-based methods could play a role instead.

8 Summary

The Bayesian approach to data analysis emphasizes full probability models and appropriate accounting for uncertainty. The approach can harness prior knowledge when it is available and provides outputs that are simple to interpret. The core related notions of "exchangeability" and "hierarchical modeling" represent powerful concepts for data mining, especially for models with huge numbers of parameters.

Prior to the emergence of Markov Chain Monte Carlo methods, computational complexity limited Bayesian data analysis to small-scale problems. Nowadays, analysts

are conducting Bayesian data analyses of considerable complexity. Nonetheless, significant barriers remain for very large scale applications and further research is required. Variational approximations are especially promising.

Religious adherence to a Bayesian or non-Bayesian perspective, often a feature of past debates in the statistical community, has become an absurdity. Bayes is not a panacea, and is certainly not always the optimal approach. In particular, for applications where accounting for uncertainty is not so important, the fully Bayesian approach may be more trouble than it is worth. Nonetheless, it is a tool of considerable flexibility and elegance, and offers many advantages for data miners.

Acknowledgements

National Science Foundation grants support Madigan's research. We thank the editor and an anonymous reviewer for helpful comments. We are grateful to David D. Lewis for stimulating conversations.

References

Barbieri, M.M and Berger, J.O. (2002). Optimal predictive model selection. Working Paper 02-02, Institute of Statistics and Decision Sciences, Duke University.

Bayarri, M.J. and Morales, J. (2000). Bayesian measures of surprise for outlier detection. `http://citeseer.nj.nec.com/bayarri00bayesian.html`.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53**, 370–418, and **54**, 296–325.

Berger, J. (2000). Bayesian analysis: a look at today and thoughts of tomorrow. Journal of the American Statistal Association, **95**, 1269–1276.

Bernardo, J.M. and A.F.M. Smith (1994). *Bayesian Theory.* John Wiley and Sons, Chichester.

Blei, D., Jordan, M., and Ng A. (2003). Latent Dirichlet models for applications in information retrieval. In: *Bayesian Statistics 7*, Oxford University Press, to appear.

Breiman, L. (2001). Statistical modeling: the two cultures (with discussion). *Statistical Science*, **16**, 199–231.

Breiman, L. and Friedman, J. (1984). Tools for large data set analysis. In: *Statistical Signal Processing*, E.J. Wegman and J.G Smith (Eds.), New York, M. Dekker, 191-197.

Cadez, I. and Smyth, P. (1999). Probabilistic clustering using hierarchical models Technical Report UCI-ICS 99-16.

Cadez, I.V., Smyth, P., Ip, E., and Mannila, H. (2001). Predictive profiles for transaction data using finite mixture models . Technical Report UCI-ICS 01-67.

Carlin, B.P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society (Series B)*, **57**, 473–484.

Carlin, B.P. and Louis, T.A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis (second edition)* Chapman and Hall, New York.

Chaloner, K. (1996). The elicitation of prior distributions. In: *Bayesian Biostatistics*, eds. D, Berry and D. Stangl, New York: Marcel Dekker, 141–156.

Cheeseman, P. and Stutz, J. (1996). Bayesian Classification (AutoClass): Theory and Results. In: *Advances in Knowledge Discovery and Data Mining*, Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, Eds. AAAI Press/MIT Press.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, **96**, 270–281.

Chipman, H.A., George, E.I., and McCulloch, R.E. (1998). Bayesian CART model search (with discussion). *Journal of the American Statistical Association*, **93**, 935–960.

Chipman, H.A., George, E.I., and McCulloch, R.E. (2001). Bayesian treed models. *Machine Learning*, to appear.

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, to appear.

Congdon, P. (2001). *Bayesian Statistical Modelling*, Wiley, London.

Consonni, G. and Veronese, P. (1995). A Bayesian method for combining results from sevaral binomial experiments. *Journal of the American Statistical Assocation*, **90**, 935–944.

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

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

Denison, D., Mallick, B., and Smith, A.F.M. (1998). A Bayesian CART algorithm. *Biometrika*, **85**, 363–377.

Denison, D.G.T., Adams, N.M., Holmes, C.C. and Hand, D.J. (2002). Bayesian partition modelling . *Computational Statistics and Data Analysis*, to appear.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society (Series B)*, **57**, 45–97.

Draper, D., Hodges, J.S., Mallows, C., and Pregibon, D. (1993). Exchangeability and data analysis (with discussion). *Journal of the Royal Statistical Society (Series A)*, **156**, 9–37.

Draper, D. and Madigan, D. (1997). The scientific value of Bayesian statistical methods. *IEEE Intelligent Systems and their Applications*, **12**, 18-21.

Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. Wiley, New York.

DuMouchel, W. (1999). Bayesian data mining in large frequency tables, with an

application to the FDA Spontaneous Reporting System. *The American Statistician*, **53**, 177-202.

DuMouchel W. and Pregibon, D. (2001). Empirical Bayes screening for multi-item associations. Proc. KDD 2001, ACM Press, San Diego, CA,

Fraley, C. and Raftery, A.E. (1998). How many clusters? Which clustering method? - Answers via Model-Based Cluster Analysis. *The Computer Journal*, **41**, 578–588.

Geisser, S. (1993). *Predictive Inference.* London: Chapman and Hall.

Gelfand, A.E. and Dey, D.K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society (Series B)*, **56**, 501-514.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis*. Chapman and Hall, London.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. (1996). *Markov chain Monte Carlo in Practice.* Chapman and Hall, London.

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711-32.

Han, C. and Carlin B.P. (2001). MCMC methods for computing Bayes factors: a comparative review. *Journal of the American Statistical Association*, to appear.

Hand D.J. and Yu K. (2001). Idiot's Bayes - not so stupid after all? *International Statistical Review*, **69**, 385–398.

Hastie, T., Tibshirani, T., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York.

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

Heckerman, D., Geiger, D., and Chickering, D.M. (1994). Learning Bayesian networks: The combination of knowledge and statistical data. In *Uncertainty in Artificial Intelligence, Proceedings of the Tenth Conference* (R. Lopez de Mantaras and D. Poole, eds.), San Francisco: Morgan Kaufmann, p.293–301.

Herbrich, R., Graepel, T., and Campbell, C. (2001). Bayes Point Machines. *Journal of Machine Learning Research*, **1**, 245–279.

Hoeting, J., Raftery, A.E., and Madigan, D. (1996). A Method for Simultaneous Variable Selection and Outlier Identification in Linear Regression *Computational Statistics and Data Analysis*, **22**, 251-270.

Hoeting, J., Madigan, D., Raftery, A.E., and Volinsky, C.T. (1999). Bayesian Model Averaging - A Tutorial (with discussion). *Statistical Science*, **14**, 382–417.

Ibrahim, J.G., Chen, M-H., and Sinha, D. (2001). *Bayesian Survival Analysis.* New York, Springer.

Jaakola, T. and Jordan, M.I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, **10**, 25–37.

Justel, A. and Pena, D. (1999). Heterogeneity and model uncertainty in Bayesian regression models. *Revista de la Real Academia de Ciencias*, **93**, 357–366.
`http://www.adi.uam.es/~ajustel/papers/jp3.pdf`

Karlin, S. and Taylor, H.M. (1975). *A First Course in Stochastic Processes (second edition).* Academic Press, New York.

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

Lewis, D.D. (1988). Naive (Bayes) at forty: The independence assumption in information retrieval. In: *ECML'98, The Tenth European Conference on Machine Learning*, 4–15.

Lewis, D.D. (2002). Personal communication.

Lie, R.T., Heuch, I., and Irgens, L.M. (1991) 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.

Lie, R.T., Heuch, I., and Irgens, L.M. (1994) Estimation of the proportion of congen-

ital malformations using double registration schemes. *Biometrics*,

Madigan, D. and Raftery, A.E. (1994). Model Selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, **89**, 1535-1546.

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

Madigan, D., Gavrin, J., and Raftery, A.E. (1995). Eliciting prior information to enhance the predictive performance of Bayesian graphical models. *Communications in Statistics - Theory and Methods*, **24**, 2271-2292.

Neal, R.M. (1996). *Bayesian Learning for Neural Networks*. Springer-Verlag.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufman.

Pole, A., West, M., and Harrison, J. (1994). *Applied Bayesian Forecasting and Time Series Analysis*. Chapman and Hall, London.

Raftery, A.E., Madigan, D., and Volinsky, C.T. (1996). Accounting for model uncertainty in survival analysis improves predictive performance. In: Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith A. F. M., (eds.), *Bayesian Statistics V*, Oxford University Press, 323-350.

Raftery, A.E., Madigan, D., and Hoeting, J. (1997). Bayesian model averaging for

linear regression models. *Journal of the American Statistical Association*, **92**, 179–191.

Richardson, S. and Green, P.J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society (Series B)*, **59**, 731–792.

Ridgeway, G. and Madigan, D. (2002). Bayesian analysis of massive datasets via particle filters. KDD-2002, to appear.

Robert, C.P., Celeux, G. and Diebolt, J. (1993). Bayesian estimation of hidden Markov chains: A stochastic implementation. *Statistics and Probability Letters*, **16**, 77-83.

Shibata, R. (1983). Asymptotic mean efficiency of a selection of regression variables. *Annals of the Institute of Mathematical Statistics*, **35**, 415–426.

Spiegelhalter, D.J. (1998). Bayesian graphical modeling: a case study in monitoring health outcomes. *Applied Statistics*, **47**, 115–133.

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

Spiegelhalter, D.J., Freedman, L.S., and parmar, M.K.B. (1994). Bayesian approaches to randomized trials (with discussion). *Journal of the Royal Statistical Society (Series A)*, **157**, 357–416.

Spiegelhalter, D.J., Thomas, A., and Best, N.G. (1999). *WinBUGS Version 1.2 User Manual*. MRC Biostatistics Unit.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society (Series B)*, to appear.

Stigler, S.M. (1990). *The History of Statistics : The Measurement of Uncertainty Before 1900*. Harvard University Press.

Volinsky, C.T., Madigan, D., Raftery, A.E., and Kronmal, R.A. (1997). Bayesian Model Averaging in Proportional Hazard Models: Predicting Strokes. *Applied Statistics*, **46**, 433-448.

Wallace, C.S. and Dowe, D.L. (2000). MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. *Statistics and Computing*, **10**,73–83.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models (second edition)*, New York, Springer.

Williams, C.K.I. (1998). Prediction with Gaussian Processes: From Linear Regression to Linear Prediction and Beyond. In: *Learning and Inference in Graphical Models*, M. I. Jordan ed., Kluwer Academic Press.

York, J., Madigan, D., Heuch, I., and Lie, R.T. (1995) Estimating a proportion of birth defects by double sampling: A Bayesian approach incorporating covariates and model uncertainty. *Applied Statistics*, **44**, 227–242.