Sparse, noisy Boolean functions

Sach Mukherjee & Terence P. Speed
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720
{sach,terry}@stat.berkeley.edu

April 20, 2007

Abstract

This paper addresses the question of making inferences regarding Boolean functions under conditions of (i) *noise*, or stochastic variation in observed data, and (ii) *sparsity*, by which we mean that the number of inputs or predictors far exceeds the arity of the underlying Boolean function. We put forward a simple probability model for such observations, and discuss model selection, parameter estimation and prediction. We present results on synthetic data and on a proteomic dataset from a study in cancer systems biology.

1 Introduction

In many applications it is natural to think of a binary output or response Y as a k-ary Boolean function of binary arguments $X_1 \dots X_k$. In this paper, we consider the problem of making inferences regarding such functions in settings characterized by (i) stochastic variation in observed data and (ii) a total number of inputs or predictors which far exceeds the arity of the underlying Boolean function. We focus on the case in which both arity k and the specific subset of predictors which are arguments to the underlying Boolean function are unknown. Thus, we are interested in inferences concerning sparse, noisy Boolean functions.

Examples of such problems are abundant in many application areas, including, among others, genetic epidemiology, data mining and systems biology. In genetic epidemiology the predictors are genetic features, such as haplotypes, while the responses are indicators of disease status. In data mining, the predictors may be, for example, a large number of observed indicators regarding customer behavior (for example, whether or not a customer purchased, or showed an interest in, each

of a large number of products), while the response of interest may be whether or not the customer will buy some new product. In systems biology, predictors may represent, for example, the activation states of proteins, while the response may be an indicator of cellular state.

In many practical settings, when confronted with a large number of potential predictors, it can be reasonable to assume that only a *small number* are relevant to the response. For example, a small number of genetic features may jointly influence disease status; equally, a small number of indicators of customer behaviour may be highly relevant to a future purchase decision. Moreover, since there are 2^k possible states of k binary arguments - and 2^{2^k} possible Boolean functions of those arguments - parsimonious models can be statistically advantageous, especially under conditions of small-to-moderate sample size.

Characterizing sparse Boolean functions from noisy data involves addressing two related problems. First, we must determine which of a possibly very large number of predictors are arguments to the underlying function; this involves selecting a subset (of unknown size) of available predictors. Second, for a putative set of k arguments, we must say something about possible k-ary Boolean functions. Statistically, we formulate these two problems as model selection and parameter estimation respectively, using a probability model introduced below.

Our work is similar in spirit to *logic regression* (Ruczinski et al., 2003; Kooperberg and Ruczinski, 2005). However, in contrast to logic regression, we focus on inferring Boolean functions rather than treating the truth value of various Boolean functions as inputs to a linear model. Our modeling approach is also very different: we do not use decision trees, but rather develop a state-dependent Binomial model. Moreover, our approach is fully Bayesian, and by making use of sparsity-promoting priors, places a clear emphasis on learning parsimonious models. Other related work on noisy Boolean functions includes Benjamini et al. (1999) and Shmulevich et al. (2002).

The remainder of this paper is organized as follows. We first introduce the key elements of our model and associated notation, and then discuss, in turn, model selection, parameter estimation and prediction. We present experimental results on synthetic data and on a proteomic dataset from a study in cancer systems biology. Finally, we discuss some of the finer points and shortcomings of our work and directions for further research.

2 Basic model and notation

2.1 Noisy Boolean functions

A k-ary *Boolean function* is a function $f : \{0,1\}^k \mapsto \{0,1\}$ which maps each of the 2^k possible states of its binary arguments $\mathbf{X} = (X_1 \ldots X_k)$ to a binary state Y. Such a function can be represented as a *truth table*:

X_1	X_2	Y
0	0	0
0	1	1
1	0	1
1	1	1

Since every distinct assignment to the right-most column of the truth table corresponds to a distinct function f, there are exactly 2^{2^k} Boolean functions of arity k.

Now, consider a function $f_\theta : \{0,1\}^k \mapsto [0,1]$, which maps each of the 2^k possible states of its arguments to the (closed) unit interval. In particular, when inputs $\mathbf{X}$ are in state $\mathbf{x}$, f_θ returns a value $\theta_\mathbf{x} = f_\theta(\mathbf{x})$ which represents the *probability* with which the output Y takes on value 1. For the moment, we do not place any restrictions on the $\theta_\mathbf{x}$'s, but we return to these parameters in the context of inference below. We call the function f_θ a *noisy Boolean function*. A noisy Boolean function can be represented by a *probabilistic truth table*:

X_1	X_2	Y
0	0	θ_{00}
0	1	θ_{01}
1	0	θ_{10}
1	1	θ_{11}

A conventional truth table can then be regarded as a special, "noise free" case of a probabilistic truth table, with parameters $\theta_\mathbf{x}$ equal to 0 or 1. It is natural to assume that if a Boolean function evaluates true for a given state $\mathbf{x}$ of its inputs, the response for a "noisy version" of the function should be true more often than false. Accordingly, if

$$\forall \mathbf{x} \cdot I_{(\frac{1}{2},1]}(f_\theta(\mathbf{x})) = f(\mathbf{x}) \tag{1}$$

where I_A is the indicator function for set A, we say that f_θ corresponds to Boolean function f. We can then construct a Boolean function f from a noisy Boolean

function f_θ by the following rule:

$$f(\mathbf{x}) = \begin{cases} 1 & \text{if } f_\theta(\mathbf{x}) > \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

This defines a (many-to-one) mapping between the space of noisy Boolean functions and the space of Boolean functions; we call this mapping Ψ and write $f = \Psi(f_\theta)$.

2.2 Probability model

Let $\mathbf{Y} = (Y_1 \dots Y_n)$, $Y_i \in \{0, 1\}$ denote binary responses and $\mathbf{X} = (\mathbf{X}_1 \dots \mathbf{X}_n)$, $\mathbf{X}_i \in \{0, 1\}^d$ corresponding d-dimensional predictors. We denote the i^{th} observation of the j^{th} predictor by X_{ij}, the i^{th} observation of predictors $A \subseteq \{1 \dots d\}$ by $\mathbf{X}_{iA}$ and the full set of n observations of predictors A by $\mathbf{X}_{\cdot A} = (\mathbf{X}_{1A} \dots \mathbf{X}_{nA})$.

Suppose Y is a noisy Boolean function of a subset $M \subseteq \{1 \dots d\}$ of predictors. The specification of this subset represents a *model*; for simplicity, we will use M to denote both the subset and the model it implies. We assume that, under model M, an observation Y_i is conditionally independent of all other predictors given $\mathbf{X}_{iM}$:

$$P(Y_i \mid \mathbf{X}_i, M) \quad = \quad P(Y_i \mid \mathbf{X}_{iM}) \tag{3}$$

Suppose the relevant predictors $\mathbf{X}_{iM}$ are in state $\mathbf{x}$. Then, $\theta_{\mathbf{x}} = f_\theta(\mathbf{x})$ is the corresponding parameter in the probabilistic truth table, and represents the probability of the event $Y_i = 1$ given the state of the predictors. In other words, $Y_i \mid \mathbf{X}_{iM} = \mathbf{x}$ is a Bernoulli random variable with success parameter $\theta_{\mathbf{x}}$:

$$P(Y_i = 1 \mid \mathbf{X}_{iM} = \mathbf{x}, \boldsymbol{\theta}) \quad = \quad \theta_{\mathbf{x}} \tag{4}$$

where $\boldsymbol{\theta}$ is a parameter vector with components $\theta_{\mathbf{x}}$.

We assume that, given the state of predictors $\mathbf{X}_{iM}$, $Y_1 \dots Y_n$ are independent and identically-distributed. Then the joint probability of the Y_i's, given $\mathbf{X}_{\cdot M}$, is a product of Binomials:

$$P(\mathbf{Y} \mid \mathbf{X}_{\cdot M}, \boldsymbol{\theta}) \quad = \quad \prod_{\mathbf{x} \in \{0,1\}^{|M|}} \text{Binomial}(\nu_{\mathbf{x}} \mid n_{\mathbf{x}}, \theta_{\mathbf{x}}) \tag{5}$$

where, $n_{\mathbf{x}} = \sum_{i:\mathbf{X}_{iM}=\mathbf{x}} 1$ is the number of observations in which predictors $\mathbf{X}_{\cdot M}$ are in state $\mathbf{x}$ and $\nu_{\mathbf{x}} = \sum_{i:\mathbf{X}_{iM}=\mathbf{x}} Y_i$ is the corresponding number of "successes" of Y_i when $\mathbf{X}_{\cdot M} = \mathbf{x}$.

3 Inference

In this Section we discuss model selection, parameter estimation and prediction using the model introduced above.

3.1 Model selection

Each model corresponds to a subset $M \subseteq \{1 \dots d\}$ of predictors. As such, there are, unconstrained, 2^d distinct models. Even if we restrict attention to Boolean functions with maximum arity k_{max}, the number of possible models is

$$\sum_{k=1}^{k_{max}} \binom{d}{k} \approx \mathcal{O}(d^{k_{max}})$$

The sheer size of model space - even under conditions of sparsity - makes model selection a central concern in inference regarding Boolean functions. Furthermore, since noisy Boolean functions can give rise to responses which depend on highly non-linear interactions *between* predictors, variable selection using marginal statistics will not, in general, be able to capture the joint explanatory power of a subset of predictors. In contrast, the state-dependent model introduced above allows us to consider all "Boolean" interactions between arguments. In this section, we exploit our probability model to develop a Bayesian approach to model selection in this setting.

3.1.1 Model posterior

From Bayes' rule, the posterior probability of a model M can be written as:

$$\begin{aligned} P(M \mid \mathbf{Y}, \mathbf{X}) &= \frac{P(\mathbf{Y} \mid \mathbf{X}, M) P(M \mid \mathbf{X})}{P(\mathbf{Y}, \mathbf{X})} \\ &= \frac{P(\mathbf{Y} \mid \mathbf{X}_{\cdot M}) P(M)}{\sum_{M \in \mathcal{M}} P(\mathbf{Y} \mid \mathbf{X}_{\cdot M}) P(M)} \end{aligned} \tag{6}$$

where $\mathcal{M}$ is the space of all possible models M.

The term $P(\mathbf{Y} \mid X_{\cdot M})$ represents the marginal likelihood of responses $Y_1 \dots Y_n$. This can be obtained by integrating out parameters $\boldsymbol{\theta}$:

$$P(\mathbf{Y} \mid X_{\cdot M}) = \int_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} P(\mathbf{Y} \mid X_{\cdot M}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \tag{7}$$

where $\boldsymbol{\Theta}$ represents the full parameter space. Now, for any Boolean function f, there exists some subset of $\boldsymbol{\Theta}$ which maps to f under mapping (2). Integrating out $\boldsymbol{\theta}$ therefore corresponds to averaging over all possible Boolean functions with arguments M.

3.1.2 Parameter prior

We first assume prior independence of parameters $\theta_{\mathbf{x}}$, such that $p(\boldsymbol{\theta}) = \prod_{\mathbf{x}} p(\theta_{\mathbf{x}})$. Then, from (5) and (7), we get:

$$P(\mathbf{Y} \mid \mathbf{X}_{\cdot M}) \;=\; \prod_{\mathbf{x}\in\{0,1\}^{|M|}} \int_0^1 \text{Binomial}(\nu_{\mathbf{x}} \mid n_{\mathbf{x}}, \theta_{\mathbf{x}}) p(\theta_{\mathbf{x}})\, \mathrm{d}\theta_{\mathbf{x}} \qquad (8)$$

In light of the relationship between parameters $\theta_{\mathbf{x}}$ and underlying Boolean functions, there are two properties we would like the parameter prior $p(\theta_{\mathbf{x}})$ to have. Firstly, given a model M corresponding to a Boolean function of arity $k = |M|$, we would like to assign equal probability to all Boolean functions possible under the model. Secondly, since the parameters $\theta_{\mathbf{x}}$ represent state-dependent success parameters for a noisy Boolean function, we would like the prior to prefer values close to zero or one. Now, for any continuous prior density symmetric about $\theta_{\mathbf{x}} = \frac{1}{2}$, $P(\theta_{\mathbf{x}} > \frac{1}{2}) = P(\theta_{\mathbf{x}} \leq \frac{1}{2}) = c$ (say). From mapping (2), the probability of a k-ary Boolean function f, given 2^k independent parameters $\theta_{\mathbf{x}}$ is:

$$\begin{aligned} P(f \mid \boldsymbol{\theta}) &= P(\Psi(f_\theta) = f \mid \boldsymbol{\theta}) \\ &= \prod_{\mathbf{x}:f(\mathbf{x})=1} P(\theta_{\mathbf{x}} > \frac{1}{2}) \;\cdot \prod_{\mathbf{x}:f(\mathbf{x})=0} P(\theta_{\mathbf{x}} \leq \frac{1}{2}) \\ &= \prod_{\mathbf{x}\in\{0,1\}^k} c \\ &= c^{2^k} \end{aligned}$$

which is constant over the space of k-ary Boolean functions. Thus, under prior parameter independence, any continuous prior density symmetric about $\theta_{\mathbf{x}} = \frac{1}{2}$ will display the first of our two desiderata. We therefore suggest a Beta prior with identical parameters α, β (for symmetry) and $\alpha, \beta < 1$ (to concentrate probability mass around 0 and 1):

$$p(\theta_{\mathbf{x}}) \;=\; \text{Beta}(\theta_{\mathbf{x}} \mid \alpha, \beta) \quad (\alpha = \beta,\, \alpha, \beta < 1) \qquad (9)$$

This gives the marginal likelihood (8) in closed form:

$$P(\mathbf{Y} \mid \mathbf{X}_{\cdot M}) \;=\; \prod_{\mathbf{x}\in\{0,1\}^{|M|}} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\nu_{\mathbf{x}}+\alpha)\Gamma(n_{\mathbf{x}}-\nu_{\mathbf{x}}+\beta)}{\Gamma(\alpha+\beta+n_{\mathbf{x}})} \qquad (10)$$

3.1.3 Sparse model prior

We use the model prior $P(M)$ to express an explicit preference for sparse models. We suggest the following prior:

$$P(M) \quad \propto \quad \kappa^{\min(0,\lambda-|M|)} \tag{11}$$

where parameter λ is a threshold on the subset size $|M|$, above which the prior begins to decay, and κ is a strength parameter. By default, we set $\kappa = e$.

An alternative to (11) would be a Poisson distribution with parameter corresponding to the prior expected model size. (Assuming a Binomial distribution for the arity of the underlying Boolean function, the Poisson arises naturally as the limiting case for a large number of predictors.) We prefer (11), because in contrast to the Poisson, it assigns equal probability to every model with $|M| \leq \lambda$, but like the Poisson decays rapidly for $|M| > \lambda$.

3.1.4 Markov chain Monte Carlo over model space

From (6), (10) and (11) we can evaluate the posterior probability of any given model up to proportionality. In smaller domains, and with a bound k_{max} on the arity of Boolean functions to be considered, we can explicitly enumerate all models M and thereby evaluate the full posterior. However, in general, the space $\mathcal{M}$ of models is much too large for such an approach, motivating the need for approximate inference. Here, we propose a Markov chain Monte Carlo sampler over model space.

Markov Chain Monte Carlo or *MCMC* represents a general class of stochastic simulation methods which are widely used in computational statistics. The basic idea of MCMC is to construct a Markov chain whose state space is the domain of the desired random quantity, and whose stationary distribution is its posterior. Then, simulating the Markov chain provides a means by which to make inferences based on the posterior distribution of interest.

In a *Metropolis-Hastings* sampler (Hastings, 1970), draws are made from a *proposal distribution* Q, which depends on current state, and then accepted or rejected in such a way as to guarantee that, asymptotically, they behave as draws from the desired target distribution. Here, we develop a MCMC sampler of the Metropolis-Hastings type for the purpose of inferring the posterior distribution (6) over models M.

In our approach, a model is equivalent to a subset M of predictor indices $\{1 \ldots d\}$. Let $\mathcal{I}(M)$ be a set comprising all subsets which can be obtained by either adding exactly one element to the set M, or by removing exactly one ele-

ment from it. That is,

$$\mathcal{I}(M) \quad = \quad \{A \cdot (|A \setminus M| = 1 \wedge M \subset A) \vee (|M \setminus A| = 1 \wedge A \subset M)\} \tag{12}$$

Then, we suggest the following proposal distribution Q:

$$Q(M'; M) = \begin{cases} \frac{1}{|\mathcal{I}(M)|} & \text{if } M' \in \mathcal{I}(M) \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

where M and M' denote current and proposed models respectively.

Then, calculate the following Hastings ratio α:

$$\alpha \quad = \quad \frac{P(M' \mid \mathbf{Y}, \mathbf{X}) Q(M; M')}{P(M \mid \mathbf{Y}, \mathbf{X}) Q(M'; M)} \tag{14}$$

From (12) and (13) we can see that the proposal distribution is symmetric, such that $Q(M'; M) = Q(M; M')$. This means the Hastings ratio is simply:

$$\alpha \quad = \quad \frac{P(M' \mid \mathbf{Y}, \mathbf{X})}{P(M \mid \mathbf{Y}, \mathbf{X})} \tag{15}$$

A proposed model M', drawn from Q, is then *accepted* with probability $\min(1, \alpha)$, and otherwise *rejected*. If accepted, M' is added to the sequence of samples drawn, and becomes the current model. Else, M is added to the sequence of samples, and remains the current model.

Since any subset of $\{1 \dots d\}$ can be reached from an arbitrary starting subset by some sequence of addition and removal steps, the proposal distribution Q gives rise to an irreducible Markov chain. Standard results (see, e.g., Robert and Casella, 2004) then guarantee convergence to the desired posterior $P(M \mid \mathbf{Y}, \mathbf{X})$. The sampler described above is summarized in Algorithm 1.

Algorithm 1 Metropolis-Hastings sampler for model selection.

(1) Initialize model $M^{(1)}$, set $t = 1$, $M \leftarrow M^{(1)}$

(2) **Propose** $M' \sim Q(M'; M)$

(3) **Accept** M' with probability $\min(1, \alpha)$, $\alpha = \frac{P(M'|\mathbf{Y},\mathbf{X})}{P(M|\mathbf{Y},\mathbf{X})}$.

(4) **Update** If M' is accepted, $M^{(t+1)} \leftarrow M'$, $M \leftarrow M^{(t+1)}$ else $M^{(t+1)} \leftarrow M$. Set $t \leftarrow t + 1$

(5) While $t < T$, repeat (2)-(4).

Importantly, during sampling, the unnormalized quantities $P(\mathbf{Y} \mid \mathbf{X}_{\cdot M'}) P(M')$ and $P(\mathbf{Y} \mid \mathbf{X}_{\cdot M}) P(M)$, which can be obtained in closed-form from (10) and (11), are sufficient for our purposes.

As shown in Algorithm 1, iterating "propose", "accept" and "update" steps gives rise to T samples $M^{(1)} \ldots M^{(T)}$. An important property of these samples is that, provided the Markov chain has converged to its stationary distribution, they provide a means by which to compute the expectation of essentially any model-dependent quantity of interest. Specifically, if $\mathbb{E}[\phi(M)]_{P(M|\mathbf{Y},\mathbf{X})}$ is the expectation, under the posterior, of a function $\phi(M)$, then

$$\hat{\mathbb{E}}[\phi(M)] \quad = \quad \frac{1}{T}\sum_{t=1}^{T}\phi(M^{(t)}) \tag{16}$$

is, by standard results, an asymptotically valid estimator of $\mathbb{E}[\phi(M)]_{P(M|\mathbf{Y},\mathbf{X})}$.

An important special case of (16), which we shall make use of below, concerns the posterior probability that a variable $j \in \{1 \ldots d\}$ is part of the underlying model M:

$$\begin{aligned} P(j \in M \mid \mathbf{Y},\mathbf{X}) &= \sum_{M \in \mathcal{M}} P(j \in M \mid M,\mathbf{Y},\mathbf{X}) P(M \mid \mathbf{Y},\mathbf{X}) \\ &= \mathbb{E}[I_M(j)]_{P(M|\mathbf{Y},\mathbf{X})} \end{aligned} \tag{17}$$

Using (16), we get an asymptotically valid estimate of $\mathbb{E}[I_M(j)]$:

$$\hat{\mathbb{E}}[I_M(j)] \quad = \quad \frac{1}{T}\sum_{t=1}^{T} I_{M^{(t)}}(j) \tag{18}$$

Finally, we note that an alternative to sampling from the posterior over models is to estimate a single, *maximum a posteriori* model M^*:

$$M^* \quad = \quad \operatorname*{argmax}_{M \in \mathcal{M}} P(M \mid \mathbf{Y},\mathbf{X})$$

This can be done using, for example, a greedy local optimization scheme in model space, with multiple, random initializations to guard against local maxima. We do not use this optimization-based approach in this paper, but note that in some settings it can be a useful and very simple approach to model selection for noisy Boolean functions.

3.2 Parameter estimation

From standard results (Gelman et al., 2004), the posterior distribution of parameter $\theta_{\mathbf{x}}$ is a Beta density:

$$p(\theta_{\mathbf{x}} \mid \mathbf{Y},\mathbf{X},M) \quad = \quad \mathrm{Beta}(\theta_{\mathbf{x}} \mid \nu_{\mathbf{x}} + \alpha, n_{\mathbf{x}} - \nu_{\mathbf{x}} + \beta) \tag{19}$$

3.3 Prediction

What is the posterior probability that a new, unseen response $Y_{(n+1)}$ will take on the value 1, given that predictors $\mathbf{X}_{(n+1)M}$ are observed in state $\mathbf{x}$? Making use of standard results (Gelman et al., 2004), it is easy to obtain the following closed-form predictive probability:

$$P(Y_{n+1} = 1 \mid \mathbf{X}_{(n+1)M} = \mathbf{x}, \mathbf{Y}, \mathbf{X}_{\cdot M})) \quad = \quad \frac{\nu_{\mathbf{x}} + \alpha}{\alpha + \beta + n_{\mathbf{x}}} \tag{20}$$

4 Results

4.1 Synthetic data

We first present an analysis of synthetic data, generated from a model in which responses depended on 4 out of d =100 predictors. Data were generated in the following manner:

(1) For $i = 1 \ldots n$ $(n = 500)$ and $j = 1 \ldots d$ $(d = 100)$, we set $X_{ij} = 1$ with probability $\frac{1}{2}$ and 0 otherwise.

(2) We specified a data-generating model by choosing a subset

$$M = \{A, B, C, D\} \subset \{1 \ldots d\}$$

of predictors, and generated responses by setting $Y_i = 1$ with probability 0.9 when $A \vee B \vee C \vee D$ (and zero otherwise) and setting $Y_i = 1$ with probability 0.1 when $\neg(A \vee B \vee C \vee D)$ (and zero otherwise).

In other words, the data-generating model was a noisy Boolean function with underlying Boolean function $f = A \vee B \vee C \vee D$ and parameters $\theta_{\mathbf{x}} = 0.9$ and $\theta_{\mathbf{x}} = 0.1$ for $f(\mathbf{x}) = 1$ and $f(\mathbf{x}) = 0$ respectively.

The model space with $|M| \leq 4$ is of size $\sim 4 \times 10^6$. We therefore eschewed exhaustive enumeration and performed model selection using MCMC, following Algorithm 1, with $T = 20,000$. To promote parsimonious models, we used sparsity prior (11), with $\lambda = 3$. Figure 1 shows average model size plotted against number of MCMC iterations: the size of sampled models converged to 3.4.

Figure 2(a) shows the posterior probabilities of the 50 most probable models encountered during sampling, ordered by probability. The model $M = \{A, B, C, D\}$ was the most probable model encountered, capturing 0.6 of the probability mass. Figure 2(b) shows posterior probabilities, calculated following (18), that each of the $d = 100$ inputs forms part of the underlying model. The four most probable

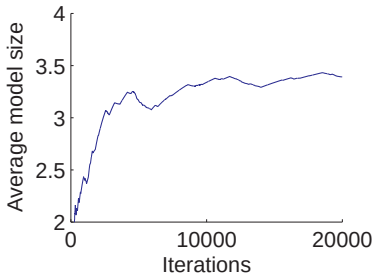
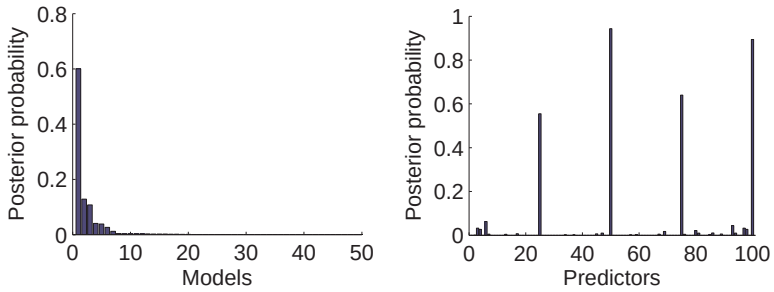


Figure 1: Synthetic data, model sparsity.

inputs were the four variables A, B, C, D in the data-generating model. In contrast, under the absolute *log odds ratios* $|\psi_j|$ between each input and the response, 3 out of the 4 correct inputs were ranked outside the top 10. (As shown in Edwards (1963), the log odds ratio is a natural measure of pairwise association for binary data.)



(a) Models (b) Individual inputs

Figure 2: Synthetic data, posterior distributions over (a) models and (b) individual inputs.

Figure 3 shows inferred posterior distributions over parameters, using the single best model encountered during sampling. The underlying Boolean function $f = A \vee B \vee C \vee D$ is false only when all its arguments are false. The posteriors over parameters clearly correspond to f: in every state except "all false", the posterior probability mass is concentrated well above $\frac{1}{2}$, but in the "all false" state (top left panel in Figure 3) probability mass is concentrated well below $\frac{1}{2}$.

Finally, we used the distribution (20) for prediction and tested the approach using leave-one-out cross-validation. This resulted in 446 correct calls out of 500, giving a leave-one-out accuracy of 89 %. (Since the noisy Boolean function gives

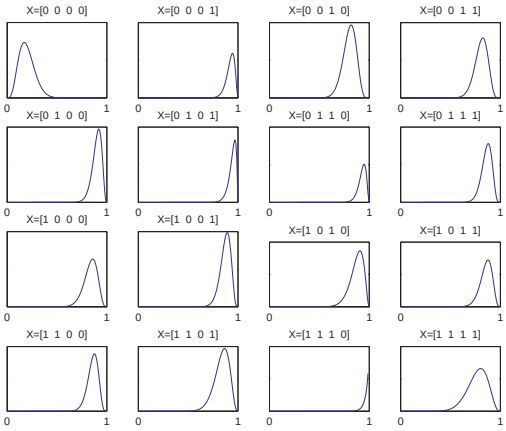


Figure 3: Synthetic data, posterior distributions over parameters.

the correct truth value with probability 0.9, expected prediction accuracy using the data-generating distribution itself would be 90%.)

4.2 Proteomic data

Our second set of empirical results concerns an analysis of proteomic data obtained from a systems-level study of breast cancer. Biological signaling systems play a central role in the biology of breast and other cancers; the data analyzed here pertain to a number of proteins involved in a key signaling system called the Epidermal Growth Factor Receptor or EGFR system. Such proteins are typically activated by a post-translational modification called phosphorylation which enables highly specific enzymatic behaviour on the part of the protein, with typically only small quantities of phosphorylated proteins required to drive downstream biochemical processes. Present/absent calls for 33 phosphorylated proteins were obtained using the KinetWorks$^{\text{TM}}$system (Kinexus Inc., Vancouver, Canada) for each of 34 breast cancer cell lines. The proteins formed a set of potential predictors. The cell lines have previously been shown (Neve et al., 2006) to reflect the diversity of primary tumors and can be usefully thought of as a sample from the space of breast tumors. The responses were a clinically important indicator called "HER2 status", which is widely used to categorize breast tumors for the purpose of targeted therapy. We sought to discover whether HER2 status could be related to the phosphorylation state of a small subset of EGFR system proteins *via* a Boolean function.

We first performed model selection using MCMC, following Algorithm 1, with $T = 10,000$ and sparsity prior (11) with $\lambda = 2$. (We chose a smaller value of λ in

this case on account of the very small sample size of $n = 34$.) Standard MCMC diagnostics showed good convergence, and the size of sampled models converged to 4.46 in this case, as shown in Figure 4.

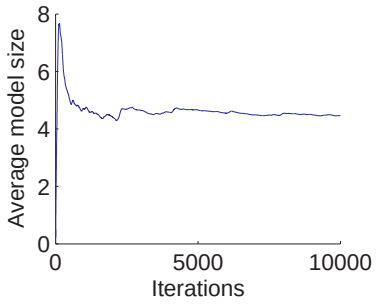
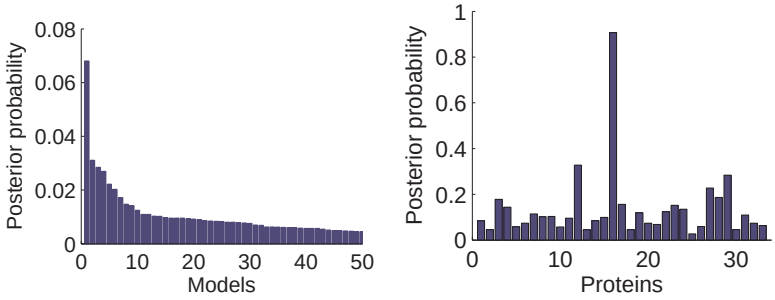


Figure 4: Proteomic data, model sparsity.

Figure 5(a) shows the posterior probabilities of the 50 most probable models encountered during sampling, ordered by probability. The single most probable model encountered had just two predictors, namely Focal Adhesion Kinase (FAK), phosphorylated on Tyrosine #576 and Insulin Receptor Substrate (IRS1), phosphorylated on Tyrosine #1179. However, this model $M = \{FAK, IRS1\}$ had a posterior probability of only 0.07. The small sample size of $n = 34$ makes the posterior distribution over models quite diffuse, such that while $M = \{FAK, IRS1\}$ has the highest posterior probability, several other models have probability on the same order of magnitude.



(a) Models (b) Individual inputs

Figure 5: Proteomic data, posterior distributions over (a) models and (b) individual inputs.

Posterior probabilities over individual inputs offer a complementary perspec-

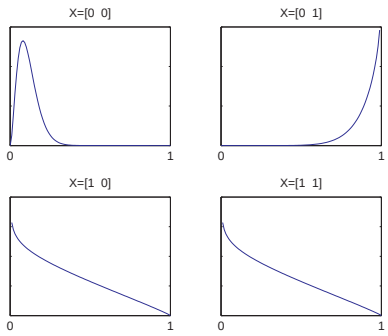


Figure 6: Proteomic data, posterior distributions over parameters.

tive to the model posterior by showing which variables appear very often in sampled models. Figure 5(b) shows posterior probabilities over individual inputs, calculated following (18). The single most probable input was IRS1, with posterior probability of 0.9. This gives us a degree of confidence in the relevance of IRS1, since it appears in most of the sampled models. The second most probable input was FAK, with posterior probability 0.33.

Figure 6 shows inferred posterior distributions over parameters, using model $M = \{FAK, IRS1\}$. The posteriors over parameters suggest that the underlying Boolean function is of the form: $\mathrm{HER2} = \mathrm{IRS1} \wedge \neg\mathrm{FAK}$. However, two of the posteriors are relatively diffuse, suggesting that the data are perhaps not sufficient to infer this rule with very high confidence.

Finally, we performed prediction using (20) and model $M = \{FAK, IRS1\}$. Leave-one-out cross-validation gave 32 out of 34 correct calls or a cross-validation accuracy of 94%.

5 Discussion and conclusions

We have presented an approach to the statistical analysis of sparse, noisy Boolean functions. We perform model selection, parameter estimation and prediction within a fully Bayesian framework, with priors on parameters designed to reflect the logical nature of underlying functions but remain agnostic otherwise, and priors on models designed to promote sparsity. Our model is simple enough to allow most quantities of interest to be computed in closed form, but general enough to describe arbitrary Boolean functions.

The size of the space of Boolean functions, and the potential complexity of such functions means that issues of over-fitting and over-confidence in inferred results are a key concern. Our use of MCMC on model space allowed us to obtain

posterior distributions over models, as well as features of models, such as the inclusion of individual predictors. These distributions allowed us to not only rank predictors and select a most probable model, but to assess our confidence in such inferences, taking into account fit to data, model complexity and number of observations. For example, in experiments on synthetic data, posteriors over models, individual predictors and parameters showed very clearly that we should have high confidence in the most probable model and the inferred Boolean function. In contrast, in our analysis of small-sample proteomic data, we found that the probability of the single best model $\{FAK, IRS1\}$ was not overwhelming, such that despite a cross-validation accuracy of 94%, there was a clear need to exercise caution in drawing definitive conclusions. However, the high posterior probabilities associated with the predictors IRS1 and, to a lesser extent, FAK, meant that we could have some confidence in their relevance. Interestingly, both proteins have been shown experimentally to exhibit interplay, *via* "crosstalk", with the wider EGFR signaling system of which HER2 forms a part (Renshaw et al., 1999; Hemi et al., 2002).

In the present paper, we specified a mapping between noisy Boolean functions and Boolean functions, but treated the characterization of a Boolean function from inferred parameters informally. In a follow-up paper, we aim to include this step in our statistical framework, and explicitly infer distributions over Boolean functions themselves.

Our results concerning HER2 status in breast cancer were largely illustrative. A more promising line of biological enquiry using noisy Boolean functions concerns the prediction of drug response from high-throughput biochemical data. We hypothesize that present/absent calls on a small number of phospho-proteins may be capable of predicting whether or not a given cancer cell line is responsive to a therapeutic agent. Sparse, Boolean prediction rules based on either proteomic analyses of this kind, or on gene expression data, could have substantive translational implications, and might also shed light on mechanisms of action or resistance.

In conclusion, sparse Boolean functions are of relevance in many areas of science and industry, and we anticipate that our work will find wide application. As noted above, our current efforts are directed towards questions in cancer systems biology, but we are also exploring applications in genetic epidemiology and data-mining.

Acknowledgements: The authors would like to thank Rich Neve and other members of Joe Gray's laboratory at Lawrence Berkeley National Laboratory for providing the proteomic dataset used in this paper. SM was supported by a Fulbright-AstraZeneca postdoctoral fellowship.

References

I. Benjamini, G. Kalai, and O. Schramm. Noise sensitivity of Boolean functions and applications to percolation. *Publications Mathématiques de l'IHÉS*, 90:5–43, 1999.

A. W. F. Edwards. The Measure of Association in a 2×2 Table. *Journal of the Royal Statistical Society. Series A (General)*, 126(1):109–114, 1963.

A. B. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. CRC Press, 2004.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97, 1970.

R. Hemi, K. Paz, N. Wertheim, A. Karasik, Y. Zick, and H. Kanety. Transactivation of ErbB2 and ErbB3 by Tumor Necrosis Factor-α and Anisomycin Leads to Impaired Insulin Signaling through Serine/Threonine Phosphorylation of IRS Proteins. *Journal of Biological Chemistry*, 277(11):8961–8969, 2002.

C. Kooperberg and I. Ruczinski. Identifying interacting SNPs using Monte Carlo logic regression. *Genetic Epidemiology*, 28(2):157–170, 2005.

R. M. Neve, K. Chin, J. Fridlyand, J. Yeh, F. L. Baehner, T. Fevr, L. Clark, N. Bayani, J.P. Coppe, F. Tong, T. P. Speed, P. T. Spellman, et al. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell*, 10(6):515–527, 2006.

M. W. Renshaw, L. S. Price, and M. A. Schwartz. Focal Adhesion Kinase Mediates the Integrin Signaling Requirement for Growth Factor Activation of MAP Kinase. *The Journal of Cell Biology*, 147(3):611–618, 1999.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

I. Ruczinski, C. Kooperberg, and M. LeBlanc. Logic regression. *Journal of Computational and Graphical Statistics*, 12(3):475–511, 2003.

I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18(2):261–274, 2002.