Bernoulli **19**(4), 2013, 1465–1483
DOI: 10.3150/12-BEJSP10

Multivariate Bernoulli distribution

BIN DAI[1], SHILIN DING[2] and GRACE WAHBA[3]

[1]*Tower Research Capital, 148 Lafayette Street, FL 12, New York, NY 10013, USA.*
E-mail: bdai@uwalumni.com
[2]*Facebook, 1601 Willow Rd, Menlo Park, CA 94025, USA. E-mail: dingsl@gmail.com*
[3]*Department of Statistics, University of Wisconsin, 1300 University Ave., Madison, WI 53706, USA.*
E-mail: wahba@stat.wisc.edu

In this paper, we consider the multivariate Bernoulli distribution as a model to estimate the structure of graphs with binary nodes. This distribution is discussed in the framework of the exponential family, and its statistical properties regarding independence of the nodes are demonstrated. Importantly the model can estimate not only the main effects and pairwise interactions among the nodes but also is capable of modeling higher order interactions, allowing for the existence of complex clique effects. We compare the multivariate Bernoulli model with existing graphical inference models – the Ising model and the multivariate Gaussian model, where only the pairwise interactions are considered. On the other hand, the multivariate Bernoulli distribution has an interesting property in that independence and uncorrelatedness of the component random variables are equivalent. Both the marginal and conditional distributions of a subset of variables in the multivariate Bernoulli distribution still follow the multivariate Bernoulli distribution. Furthermore, the multivariate Bernoulli logistic model is developed under generalized linear model theory by utilizing the canonical link function in order to include covariate information on the nodes, edges and cliques. We also consider variable selection techniques such as LASSO in the logistic model to impose sparsity structure on the graph. Finally, we discuss extending the smoothing spline ANOVA approach to the multivariate Bernoulli logistic model to enable estimation of non-linear effects of the predictor variables.

Keywords: Bernoulli distribution; generalized linear models; LASSO; smoothing spline

1. Introduction

Undirected graphical models have been proved to be useful in a variety of applications in statistical machine learning. Statisticians and computer scientists devoted resources to studies in graphs with nodes representing both continuous and discrete variables. Such models consider a graph $G = (V, E)$, whose nodes set V represents K random variables $Y_1, Y_2, \ldots, Y_K$ connected or disconnected defined by the undirected edges set E. This formulation allows pairwise relationships among the nodes to be described in terms of edges, which in statistics are defined as correlations. The graph structure can thus be determined under the independence assumptions on the random variables. Specifically, variables Y_i and Y_j are conditionally independent given all other variables if the associated nodes are not linked by an edge. Two important types of graphical models are the Gaussian model, where the K variables are assumed to follow a joint multivariate Gaussian distribution, and the Markov model, which captures the relationships between categorical variables.

However, the assumption that only the pairwise correlations among the variables are considered may not be sufficient for real applications. When the joint distribution of the nodes is

1350-7265 © 2013 ISI/BS

multivariate Gaussian, the graph structure can be directly inferred from the inverse of the covariance matrix of the random variables and in recent years a large body of literature has emerged in this area for high-dimensional data. Researchers mainly focus on different sparse structure of the graphs or, in other words, the covariance matrix for high-dimensional observations. For example, [11] proposes a consistent approach based on LASSO from [16] to model the sparsity of the graph. Due to the fact that the Gaussian distribution can be determined by the means and covariance matrix, it is valid to consider only the pairwise correlations, but this may not true for some other distributions. The multivariate Bernoulli distribution discussed in [20], which will be studied in Section 3, has a probability density function involving terms representing third and higher order moments of the random variables, which is also referred to as clique effects. To alleviate the complexity of the graph, the so-called Ising model borrowed from physics gained popularity in the machine learning literature. [19] introduces several important discrete graphical models including the Ising model and [1] discussed a framework to infer sparse graph structure with both Gaussian and binary variables. In this paper, higher than second interactions among a group of binary random variables are studied in detail. The multivariate Bernoulli model is equivalent to Ising model and other undirected graphical model with binary nodes, which has been used in the machine learning community for various applications. It can be extended to include k-node cliques by adding monomials of up to k orders [19]. The Ising model assumes the nodes taking values in $\{-1, 1\}$, which makes the interpretation of the interactions different form the multivariate Bernoulli model. The literature related to structure selection of Ising models and the applications include but are not limited to [13] and [22].

What's more, in some real applications, people are not only interested in the graph structure but also want to include predictor variables that potentially have influence on the graph structure. [6] considers a multivariate Bernoulli model which uses a smoothing spline ANOVA model to replace the linear predictor [10] for main effects on the nodes, but set the second and higher order interactions between the nodes as constants. Higher order outcomes with hierarchical structure assumptions on the graph involving predictor variables are studied in [4].

This paper aims at building a unified framework of a generalized linear model for the multivariate Bernoulli distribution which includes both higher order interactions among the nodes and covariate information. The remainder is organized as follows. Section 2 starts from the simplest multivariate Bernoulli distribution, the so-called bivariate Bernoulli distribution, where there are only two nodes in the graph. The mathematical formulation and statistical properties of the multivariate Bernoulli distribution are addressed in Section 3. Section 4 serves to get a better understanding of the differences and similarities of the multivariate Bernoulli distribution with the Ising and multivariate Gaussian models. Section 5 extends the model to include covariate information on the nodes, edges and cliques, and discusses parameter estimation, optimization and associated problems in the resulting multivariate Bernoulli logistic model. Finally, Section 6 provides conclusion of the paper and some proofs are deferred to Appendix.

2. Bivariate Bernoulli distribution

To start from the simplest case, we extend the widely used univariate Bernoulli distribution to two dimensions in this section and the more complicated multivariate Bernoulli distribution is explored in Section 3. The Bernoulli random variable Y, is one with binary outcomes chosen from {0, 1} and its probability density function is

$$f_Y(y) = p^y(1-p)^{1-y}.$$

Next, consider bivariate Bernoulli random vector (Y_1, Y_2), which takes values from $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$ in the Cartesian product space $\{0,1\}^2 = \{0,1\} \times \{0,1\}$. Denote $p_{ij} = P(Y_1 = i, Y_2 = j)$, $i, j = 0, 1$, then its probability density function can be written as

$$\begin{aligned} P(Y = y) &= p(y_1, y_2) \\ &= p_{11}^{y_1 y_2} p_{10}^{y_1(1-y_2)} p_{01}^{(1-y_1)y_2} p_{00}^{(1-y_1)(1-y_2)} \qquad (2.1)\\ &= \exp\left\{\log(p_{00}) + y_1 \log\left(\frac{p_{10}}{p_{00}}\right) + y_2 \log\left(\frac{p_{01}}{p_{00}}\right) + y_1 y_2 \log\left(\frac{p_{11}p_{00}}{p_{10}p_{01}}\right)\right\}, \end{aligned}$$

where the side condition $p_{00} + p_{10} + p_{01} + p_{11} = 1$ holds to ensure it is a valid probability density function.

To simplify the notation, define the natural parameters f's from general parameters as follows:

$$f^1 = \log\left(\frac{p_{10}}{p_{00}}\right), \tag{2.2}$$

$$f^2 = \log\left(\frac{p_{01}}{p_{00}}\right), \tag{2.3}$$

$$f^{12} = \log\left(\frac{p_{11}p_{00}}{p_{10}p_{01}}\right), \tag{2.4}$$

and it is not hard to verify the inverse of the above formula

$$p_{00} = \frac{1}{1 + \exp(f^1) + \exp(f^2) + \exp(f^1 + f^2 + f^{12})}, \tag{2.5}$$

$$p_{10} = \frac{\exp(f^1)}{1 + \exp(f^1) + \exp(f^2) + \exp(f^1 + f^2 + f^{12})}, \tag{2.6}$$

$$p_{01} = \frac{\exp(f^2)}{1 + \exp(f^1) + \exp(f^2) + \exp(f^1 + f^2 + f^{12})}, \tag{2.7}$$

$$p_{11} = \frac{\exp(f^1 + f^2 + f^{12})}{1 + \exp(f^1) + \exp(f^2) + \exp(f^1 + f^2 + f^{12})}. \tag{2.8}$$

Here the original density function (2.1) can be viewed as a member of the exponential family, and represented in a log-linear formulation as:

$$P(Y = y) = \exp\{\log(p_{00}) + y_1 f^1 + y_2 f^2 + y_1 y_2 f^{12}\}. \tag{2.9}$$

Consider the marginal and conditional distribution of Y_1 in the random vector (Y_1, Y_2), we have

Proposition 2.1. *The marginal distribution of Y_1 in a bivariate Bernoulli vector (Y_1, Y_2) following density function (2.1) is univariate Bernoulli with density*

$$P(Y_1 = y_1) = (p_{10} + p_{11})^{y_1}(p_{00} + p_{01})^{(1-y_1)}. \tag{2.10}$$

What's more, the conditional distribution of Y_1 given Y_2 is also univariate Bernoulli with density

$$P(Y_1 = y_1|Y_2 = y_2) = \left(\frac{p(1, y_2)}{p(1, y_2) + p(0, y_2)}\right)^{y_1}\left(\frac{p(0, y_2)}{p(1, y_2) + p(0, y_2)}\right)^{1-y_1}. \tag{2.11}$$

The proposition implies that the bivariate Bernoulli distribution is similar to the bivariate Gaussian distribution, in that both the marginal and conditional distributions are still Bernoulli distributed. On the other hand, it is also important to know under what conditions the two random variables Y_1 and Y_2 are independent.

Lemma 2.1. *The components of the bivariate Bernoulli random vector (Y_1, Y_2) are independent if and only if f^{12} in (2.9) and defined in (2.4) is zero.*

The Lemma 2.1 is a special case for Theorem 3.1 in Section 3, and the proof is attached in Appendix. It is not hard to see from the log-linear formulation (2.9) that when $f^{12} = 0$, the probability density function of the bivariate Bernoulli is separable in y_1 and y_2 so the lemma holds. In addition, a simple calculation of covariance between Y_1 and Y_2 gives

$$\begin{aligned} \operatorname{cov}(Y_1, Y_2) &= E\big[Y_1 - (p_{11} + p_{10})\big]\big[Y_2 - (p_{11} + p_{01})\big] \\ &= p_{11}p_{00} - p_{01}p_{10}, \end{aligned} \tag{2.12}$$

and using (2.4), the disappearance of f^{12} indicates that the correlation between Y_1 and Y_2 is null. When dealing with the multivariate Gaussian distribution, the uncorrelated random variables are independent as well and Section 3 below shows uncorrelatedness and independence is also equivalent for the multivariate Bernoulli distribution.

The importance of Lemma 2.1 was explored in [20] where it was referred to as Proposition 2.4.1. The importance of f^{12} (denoted as *u-terms*) is discussed and called *cross-product ratio* between Y_1 and Y_2. The same quantity is actually *log odds* described for the univariate case in [10] and for the multivariate case in [9].

3. Formulation and statistical properties

3.1. Probability density function

As discussed in Section 2, the two dimensional Bernoulli distribution possesses good properties analogous to the Gaussian distribution. This section is to extend it to high-dimensions and construct the so-called multivariate Bernoulli distribution.

Let $Y = (Y_1, Y_2, \dots, Y_K)$ be a K-dimensional random vector of possibly correlated Bernoulli random variables (binary outcomes) and let $y = (y_1, \dots, y_K)$ be a realization of Y. The most general form $p(y_1, \dots, y_K)$ of the joint probability density is

$$\begin{aligned} P(Y_1 = y_1, Y_2 = y_2, \dots, Y_K = y_K) &= p(y_1, y_2, \dots, y_K) \\ &= p(0, 0, \dots, 0)^{[\prod_{j=1}^K (1-y_j)]} \\ &\quad \times p(1, 0, \dots, 0)^{[y_1 \prod_{j=2}^K (1-y_j)]} \\ &\quad \times p(0, 1, \dots, 0)^{[(1-y_1) y_2 \prod_{j=3}^K (1-y_j)]} \dots \\ &\quad \times p(1, 1, \dots, 1)^{[\prod_{j=1}^K y_j]}, \end{aligned}$$

or in short

$$p(y) = p_{0,0,\dots,0}^{[\prod_{j=1}^K (1-y_j)]} p_{1,0,\dots,0}^{[y_1 \prod_{j=2}^K (1-y_j)]} p_{0,1,\dots,0}^{[(1-y_1) y_2 \prod_{j=3}^K (1-y_j)]} \cdots p_{1,1,\dots,1}^{[\prod_{j=1}^K y_j]}. \tag{3.1}$$

To simplify the notation, denote the quantity S to be

$$S^{j_1 j_2 \cdots j_r} = \sum_{1 \le s \le r} f^{j_s} + \sum_{1 \le s < t \le r} f^{j_s j_t} + \cdots + f^{j_1 j_2 \cdots j_r}, \tag{3.2}$$

and in the bivariate Bernoulli case $S^{12} = f^1 + f^2 + f^{12}$. To eliminate the product in the tedious exponent of (3.1), define the interaction function B

$$B^{j_1 j_2 \cdots j_r}(y) = y_{j_1} y_{j_2} \cdots y_{j_r}, \tag{3.3}$$

so correspondingly in the bivariate Bernoulli distribution for the realization (y_1, y_2) of random vector (Y_1, Y_2), the interaction function of order 2 is $B^{12}(y) = y_1 y_2$. This is the only available order two interaction for the bivariate case. In general, there are $\binom{K}{2} = \frac{K(K-1)}{2}$ different second interactions among the binary components of the multivariate Bernoulli random vector.

The log-linear formulation of the multivariate Bernoulli distribution induced from (3.1) is

$$\begin{aligned} l(y, \mathbf{f}) &= -\log\bigl[p(y)\bigr] \\ &= -\left[\sum_{r=1}^K \left(\sum_{1 \le j_1 < j_2 < \cdots < j_r \le K} f^{j_1 j_2 \cdots j_r} B^{j_1 j_2 \cdots j_r}(y) \right) - b(\mathbf{f}) \right], \end{aligned} \tag{3.4}$$

where $\mathbf{f} = (f^1, f^2, \dots, f^{12\cdots K})^T$ is the vector of the natural parameters for multivariate Bernoulli, and the normalizing factor $b(\mathbf{f})$ is defined as

$$b(\mathbf{f}) = \log \sum_{r=1}^K \left[1 + \left(\sum_{1 \le j_1 < j_2 < \cdots < j_r \le K} \exp\bigl[S^{j_1 j_2 \cdots j_r}\bigr] \right) \right]. \tag{3.5}$$

As a member of the exponential distribution family, the multivariate Bernoulli distribution has the fundamental ‘link’ between the natural and general parameters.

Lemma 3.1 (Parameter transformation). *For the multivariate Bernoulli model, the general parameters and natural parameters have the following relationship.*

$$\exp\bigl(f^{j_1 j_2 \cdots j_r}\bigr) = \frac{\prod p(\textit{even \# zeros among } j_1, j_2, \ldots, j_r \textit{ components and other components are all zero})}{\prod p(\textit{odd \# zeros among } j_1, j_2, \ldots, j_r \textit{ components and other components are all zero})},$$

where # refers to the number of zeros among the superscript $y_{j_1} \cdots y_{j_r}$ of f. In addition,

$$\exp\bigl(S^{j_1 j_2 \cdots j_r}\bigr) = \frac{p(j_1, j_2, \ldots, j_r \textit{ positions are one, others are zero})}{p(0, 0, \ldots, 0)} \tag{3.6}$$

and conversely the general parameters can be represented by the natural parameters

$$p(j_1, j_2, \ldots, j_r \textit{ positions are one, others are zero}) = \frac{\exp(S^{j_1 j_2 \cdots j_r})}{\exp(b(\mathbf{f}))}. \tag{3.7}$$

Based on the log-linear formulation (3.4) and the fact that the multivariate Bernoulli distribution is a member of the exponential family, the interactions functions $B^{j_1 j_2 \cdots j_r}(y)$ for all combinations $j_1 j_2 \cdots j_r$ define the sufficient statistics. In addition, the log-partition function $b(\mathbf{f})$ as in (3.5) is useful to determine the expectation and variance of the sufficient statistics to be addressed in later sections.

3.2. Independence, marginal and conditional distributions

One of the most important statistical properties for the multivariate Gaussian distribution is the equivalence of independence and uncorrelatedness. As a natural multivariate extension of the univariate Bernoulli distribution, it is of great interest to explore independence among components of the multivariate Bernoulli distribution and it is the topic for this section.

The independence of components of a random vector is determined by separability of coordinates in its probability density function and it is hard to get directly from (3.1). However, based on the relationship between the natural parameters and the outcome in the log-linear formulation (3.4), the independence theorem of the distribution can be derived as follows with proof deferred to Appendix.

Theorem 3.1 (Independence of Bernoulli outcomes). *For the multivariate Bernoulli distribution, the random vector $Y = (Y_1, \ldots, Y_K)$ is independent element-wise if and only if*

$$f^{j_1 j_2 \cdots j_r} = 0 \qquad \forall 1 \le j_1 < j_2 < \cdots < j_r \le K, r \ge 2. \tag{3.8}$$

In addition, the condition in equation (3.8) *can be equivalently written as*

$$S^{j_1 j_2 \cdots j_r} = \sum_{k=1}^{r} f^{j_k} \qquad \forall r \geq 2. \tag{3.9}$$

The importance of the theorem is to link the independence of components of a random vector following the multivariate Bernoulli distribution to the natural parameters. Notice that to ensure all the single random variable to be independent of all the others is a strong assertion and in graphical models, researchers are more interested in the independence of two groups of nodes, so we have the following theorem:

Theorem 3.2 (Independence of groups). *For random vector* $Y = (Y_1, \ldots, Y_K)$ *following the multivariate Bernoulli distribution, without of loss of generality, suppose two blocks of nodes* $Y' = (Y_1, Y_2, \ldots, Y_r)$, $Y'' = (Y_{r+1}, Y_{r+2}, \ldots, Y_s)$ *with* $1 \leq r < s \leq K$, *and denote index set* $\tau_1 = \{1, 2, \ldots, r\}$ *and* $\tau_2 = \{r+1, r+2, \ldots, s\}$. *Then* Y' *and* Y'' *are independent if and only if*

$$f^{\tau} = 0 \qquad \forall \tau \cap \tau_1 \neq \emptyset \text{ and } \tau \cap \tau_2 \neq \emptyset. \tag{3.10}$$

The proof of Theorem 3.2 is also deferred to Appendix. The theorem delivers the message that the two groups of binary nodes in a graph are independent if all the natural parameters f's corresponding to the index sets that include indices from both groups disappear.

Furthermore, analogous to the multivariate Gaussian distribution, researchers are interested in statistical distributions of marginal and conditional distributions for the multivariate Bernoulli distribution. Likewise, the multivariate Bernoulli distribution maintains the good property that both the marginal and conditional distributions are still multivariate Bernoulli as stated in the following proposition.

Proposition 3.1. *The marginal distribution of the random vector* $(Y_1, \ldots, Y_K)$ *which follows multivariate Bernoulli distribution with density function* (3.1) *to any order is still a* multivariate Bernoulli *with density*

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_r = y_r) = \sum_{y_{r+1}} \cdots \sum_{y_K} p(y_1, \ldots, y_K) \tag{3.11}$$

for some $r < K$.

What's more, the conditional distribution of $(Y_1, Y_2, \ldots, Y_r)$ *given the rest is also* multivariate Bernoulli *with density*

$$P(Y_1 = y_1, \ldots, Y_r = y_r | Y_{r+1} = y_{r+1}, \ldots, Y_K = y_K) = \frac{p(y_1, \ldots, y_K)}{p(y_{r+1}, \ldots, y_K)}. \tag{3.12}$$

3.3. Moment generating functions

The moment generating function for the multivariate Bernoulli distribution is useful when dealing with moments and proof of Theorem 3.1.

$$\begin{aligned}\psi(\mu_1,\mu_2,\dots,\mu_K) &= E\bigl[\exp(\mu_1 Y_1+\mu_2 Y_2+\cdots+\mu_K Y_K)\bigr]\\ &= p_{00\cdots 0}e^0 + p_{10\cdots 0}e^{\mu_1}+\cdots+p_{11\cdots 1}e^{\mu_1+\mu_2+\cdots+\mu_K}\\ &= \sum_{r=1}^{K}\sum_{j_1\le j_2\le\cdots\le j_r}\frac{\exp[S^{j_1 j_2\cdots j_r}]}{\exp[b(\mathbf{f})]}\exp\Biggl[\sum_{k=1}^{r}\mu_{j_k}\Biggr].\end{aligned} \tag{3.13}$$

Hence, from the formula the moment generating function is solely determined by the S functions, which are the transformation of the natural parameters f's.

3.4. Gradient and Hessian

As a member of the exponential family, the gradient and Hessian (Fisher information) are the mean and covariance of the random vector $(Y_1, Y_2, \dots, Y_K)$. Therefore, they are important in statistics but also crucial for model inference when the proper optimization problem is established. To examine the formulation of gradient and Hessian for the logarithm of the multivariate Bernoulli distribution (3.1), let us define some notations.

Denote $\mathcal{T}$ to be the set of all possible superscripts of the f's including the null superscript with $f^{\emptyset}=0$, so it has 2^K elements. In other words, $\mathcal{T}$ is the power set of indices $\{1,2,\dots,K\}$. Let $|\cdot|$ be the cardinality of a set then $|\mathcal{T}|=2^K$. We can define the relation subset $\subset$ for $\tau_1,\tau_2\in\mathcal{T}$ as follows.

Definition 3.1. *For any two superscripts $\tau_1=\{j_1,j_2,\dots,j_r\}$ such that $\tau_1\in\mathcal{T}$ and $\tau_2=\{k_1,k_2,\dots,k_s\}$ with $\tau_2\in\mathcal{T}$ and $r\le s$, we say that $\tau_1\subseteq\tau_2$ if for any $j\in\tau_1$, there is a $k\in\tau_2$ such that $j=k$.*

Based on the definition, the S's in (3.2) can be reformulated as

$$S^{\tau}=\sum_{\tau_0\subseteq\tau} f^{\tau_0}, \tag{3.14}$$

specifically, $S^{\emptyset}=0$. Consider the gradient of the log-linear form (3.4) with respect to the f's, for any $\tau\in\mathcal{T}$,

$$\begin{aligned}\frac{\partial l(y,\mathbf{f})}{\partial f^{\tau}} &= -B^{\tau}(y)+\frac{\partial b(\mathbf{f})}{\partial f^{\tau}}\\ &= -B^{\tau}(y)+\frac{\sum_{\tau_0\supseteq\tau}\exp[S^{\tau_0}]}{b(\mathbf{f})}.\end{aligned} \tag{3.15}$$

The derivation of partial derivative of b with respect to f^τ in (3.15) is

$$\begin{aligned}\frac{\partial b(\mathbf{f})}{\partial f^\tau} &= \frac{1}{\exp[b(\mathbf{f})]}\cdot\frac{\partial\exp[b(\mathbf{f})]}{\partial f^\tau}\\ &= \frac{1}{\exp[b(\mathbf{f})]}\cdot\frac{\partial\sum_{\tau_0\in\mathcal{T}}\exp[S^{\tau_0}]}{\partial f^\tau}\\ &= \frac{\sum_{\tau_0\supseteq\tau}\exp[S^{\tau_0}]}{\exp[b(\mathbf{f})]}\\ &= E\bigl[B^\tau(y)\bigr],\end{aligned}\tag{3.16}$$

and the result can also be derived from the moment generating function (3.13) by taking derivatives with respect to the μ's.

A simple example of (3.15) in the bivariate Bernoulli distribution (2.9) is

$$\frac{\partial l(y,\mathbf{f})}{\partial f^1} = -y_1 + \frac{\exp(f^1)+\exp(S^{12})}{b(\mathbf{f})}.$$

Further, the general formula for the second order derivative of (3.4) with respect to any two natural parameters f^{τ_1} and f^{τ_2} is

$$\begin{aligned}\frac{\partial^2 l(y,f)}{\partial f^{\tau_1}\partial f^{\tau_2}} &= \frac{\partial^2 b(\mathbf{f})}{\partial f^{\tau_1}\partial f^{\tau_2}}\\ &= \frac{\partial}{\partial f^{\tau_1}}\left(\frac{\sum_{\tau_0\supseteq\tau_2}\exp[S^{\tau_0}]}{\exp[b(\mathbf{f})]}\right)\\ &= \frac{\sum_{\tau_0\supseteq\tau_1,\tau_0\supseteq\tau_2}\exp[S^{\tau_0}]\exp[b(\mathbf{f})]-\sum_{\tau_0\supseteq\tau_1}\exp[S^{\tau_0}]\sum_{\tau_0\supseteq\tau_2}\exp[S^{\tau_0}]}{\exp[2b(\mathbf{f})]}\\ &= \operatorname{cov}\bigl(B^{\tau_1}(y),B^{\tau_2}(y)\bigr).\end{aligned}\tag{3.17}$$

In the bivariate Bernoulli distribution,

$$\frac{\partial^2 l(y,f)}{\partial f^1\partial f^2} = \frac{\exp[S^{12}]\exp[b(\mathbf{f})]-(\exp[f^1]+\exp[S^{12}])(\exp[f^2]+\exp[S^{12}])}{\exp[2b(\mathbf{f})]}.$$

4. The Ising and the multivariate Gaussian models

As mentioned in Section 1, the Ising and the multivariate Gaussian distributions are two main tools to study undirected graphical models, and this section is to compare the multivariate Bernoulli model introduced in Section 3 with these two popular models.

4.1. The Ising model

The Ising model, which originated from [8], becomes popular when the graph structure is of interest with nodes taking binary values. The log-linear density of the random vector $(Y_1, \ldots, Y_K)$ is

$$\log\bigl[f(Y_1, \ldots, Y_K)\bigr] = \sum_{j=1}^{K} \theta_{j,j} Y_j + \sum_{1 \le j < j' \le K} \theta_{j,j'} Y_j Y_{j'} - \log\bigl[Z(\Theta)\bigr], \tag{4.1}$$

where $\Theta = (\theta_{j,j'})_{K \times K}$ is a symmetric matrix specifying the network structure, but it is not necessarily positive semi-definite. The log-partition function $Z(\Theta)$ is defined as

$$Z(\Theta) = \sum_{Y_j \in \{0,1\}, 1 \le j \le K} \exp\left(\sum_{j=1}^{K} \theta_{j,j} Y_j + \sum_{1 \le j < j' \le K} \theta_{j,j'} Y_j Y_{j'} \right), \tag{4.2}$$

and notice that it is not related to Y_j due to the summation over all possible values of Y_j for $j = 1, 2, \ldots, K$.

It is not hard to see that the multivariate Bernoulli is an extension of the Ising model, which assumes all $S^\tau = 0$ for any τ such that $|\tau| > 2$ and $\theta_{j,j'} = S^{jj'}$. In other words, in the Ising model, only pairwise interactions are considered. [13] pointed out that the higher order interactions, which is referred to as clique effects in this paper, can be converted to pairwise ones through the introduction of additional variables and thus retain the Markovian structure of the network defined in [19].

4.2. Multivariate Gaussian model

When continuous nodes are considered in a graphical model, the multivariate Gaussian distribution is important since, similar to the Ising model, it only considers interactions up to order two. The log-linear formulation is

$$\log\bigl[f(Y_1, \ldots, Y_K)\bigr] = \bigl(-\tfrac{1}{2}(Y - \mu)^T \Sigma (Y - \mu)\bigr) - \log\bigl[Z(\Sigma)\bigr], \tag{4.3}$$

where $Z(\Sigma)$ is the normalizing factor which only depends on the covariance matrix Σ.

4.3. Comparison of different graphical models

The multivariate Bernoulli (3.4), Ising (4.1) and multivariate Gaussian (4.3) are three different kinds of graphical models and they share many similarities

1. All of them are members of the exponential family.
2. Uncorrelatedness and independence are equivalent.
3. Conditional and marginal distributions maintain the same structure.

Table 1. The number of parameters in the multivariate Bernoulli, the Ising and the multivariate Gaussian models

Graph dimension	Multivariate Bernoulli	Ising	Multivariate Gaussian
1	1	1	2
2	3	3	5
3	7	6	9
...	...	...	...
K	$2^K - 1$	$\frac{K(K+1)}{2}$	$K + \frac{K(K+1)}{2}$

However, some differences do exist. the multivariate Bernoulli and the Ising models both serve as tools to model graph with binary nodes, and are certainly different from the multivariate Gaussian model which formulates continuous variables. In addition, the multivariate Bernoulli specifies clique effects among nodes whereas the Ising model simplifies to deal with only pairwise interactions and the multivariate Gaussian essentially is uniquely determined by its mean and covariance structure, which is also based on first and second order moments. Table 1 illustrates the number of parameters needed to uniquely determine the distribution for these models as the number of nodes K in the graph increases.

5. Multivariate Bernoulli logistic models

5.1. Generalized linear model

As discussed in Section 3, the multivariate Bernoulli distribution is a member of the exponential family and as a result, the generalized linear model theory in [10] applies. The natural parameters (f's) in Lemma 3.1 can be formulated as a linear predictor in [10] such that for any $\tau \in \mathcal{T}$ with $\mathcal{T} = \{1, 2, \dots, K\}$

$$f^\tau(x) = c_0^\tau + c_1^\tau x_1 + \cdots + c_p^\tau x_p, \tag{5.1}$$

where the vector $c^\tau = (c_0^\tau, \dots, c_p^\tau)$ for $\tau \in \mathcal{T}$ is the coefficient vector to be estimated and $x = (x_1, x_2, \dots, x_p)$ is the observed covariate. Here p is the number of variables and there are $2^K - 1$ coefficient vectors to be estimated so in total $p \times (2^K - 1)$ unknown parameters. Equation (5.1) is built on the canonical link where natural parameters are directly modeled as linear predictors, but other links are possible and valid as well.

When there are n samples observed from a real data set with outcomes denoted as $y(i) = (y_1(i), \dots, y_K(i))$ and predictor variables $x(i) = (x_1(i), \dots, x_p(i))$, the negative log likelihood for the generalized linear model of the multivariate Bernoulli distribution is

$$l\bigl(y, \mathbf{f}(x)\bigr) = \sum_{i=1}^{n} \left[-\sum_{\tau \in \mathcal{T}} f^\tau\bigl(x(i)\bigr) B^\tau\bigl(y(i)\bigr) + b\bigl(\mathbf{f}(x)\bigr) \right], \tag{5.2}$$

where, similar to (3.5) the log partition function b is

$$b\big(\mathbf{f}(x)\big) = \log\Bigg[1 + \sum_{\tau\in\mathcal{T}} \exp\big[S^\tau\big(x(i)\big)\big]\Bigg].$$

When dealing with the univariate Bernoulli distribution using formula (5.2), the resulting generalized linear model corresponding to the multivariate Bernoulli model is the same for logistic regression. Thus the model is referred to as the multivariate Bernoulli logistic model in this paper.

5.2. Gradient and Hessian

To optimize the negative log likelihood function (5.1) with respect to the coefficient vector c^τ, the efficient and popular iterative re-weighted least squares algorithm mentioned in [10] can be implemented. Nevertheless, the gradient vector and Hessian matrix (Fisher Information) with respect to the coefficients c^τ are still required.

Consider any $\tau \in \mathcal{T}$, the first derivative with respect to c_j^τ in the negative log likelihood (5.2) of the multivariate Bernoulli logistic model, according to (3.15) and ignoring index i, is

$$\begin{aligned}\frac{\partial l(y,f)}{\partial c_j^\tau} &= \frac{\partial l(y,f)}{\partial f^\tau}\frac{\partial f^\tau}{\partial c_j^\tau} \\ &= \sum_{i=1}^n \Bigg[-B^\tau(y) + \frac{\sum_{\tau_0\supseteq\tau}\exp[S^{\tau_0}(x)]}{\exp[b(\mathbf{f}(x))]}\Bigg]x_j.\end{aligned} \tag{5.3}$$

Further, the second derivative for any two coefficients $c_j^{\tau_1}$ and $c_k^{\tau_2}$ is

$$\begin{aligned}\frac{\partial^2 l(y,f)}{\partial c_j^{\tau_1}\partial c_k^{\tau_2}} &= \frac{\partial}{\partial c_j^{\tau_1}}\left(\frac{\partial l(y,f)}{\partial f^{\tau_2}}\frac{\partial f^{\tau_2}}{\partial c_k^{\tau_2}}\right) \\ &= \frac{\partial f^{\tau_1}}{\partial c_j^{\tau_1}}\frac{\partial^2 l(y,f)}{\partial f^{\tau_1}\partial f^{\tau_2}}\frac{\partial f^{\tau_2}}{\partial c_k^{\tau_2}} \\ &= \sum_{i=1}^n \frac{\partial^2 l(y,f)}{\partial f^{\tau_1}\partial f^{\tau_2}}x_j x_k \\ &= \frac{\sum_{\tau_0\supseteq\tau_1,\tau_0\supseteq\tau_2}\exp[S^{\tau_0}(x)]}{\exp[b(f(x))]}x_j x_k - \frac{\sum_{\tau_0\supseteq\tau_1}\exp[S^{\tau_0}(x)]\sum_{\tau_0\supseteq\tau_2}\exp[S^{\tau_0}(x)]}{\exp[2b(f(x))]}x_j x_k.\end{aligned} \tag{5.4}$$

5.3. Parameters estimation and optimization

With gradient (5.3) and Hessian (5.4) at hand, the minimization of the negative log likelihood (5.2) with respect to the coefficients c^τ can be solved with Newton–Raphson or the Fisher's scoring algorithm (iterative re-weighted least squares) when the Hessian is replaced by the Fisher information matrix. Therefore, in every iteration, the new step size for current estimate $\hat{c}^{(s)}$ is computed as

$$\triangle c = -\left(\frac{\partial^2 l(y,f)}{\partial c_j^{\tau_1} \partial c_k^{\tau_2}}\bigg|_{c=\hat{c}^{(s)}}\right)^{-1} \cdot \left(\frac{\partial l(y,f)}{\partial c_j^{\tau}}\bigg|_{c=\hat{c}^{(s)}}\right). \tag{5.5}$$

The process continues until the convergence criterion is met.

5.4. Variable selection

Variable selection is important in modern statistical inference. It is also crucial to select only the significant variables to determine the structure of the graph for better model identification and prediction accuracy. The pioneering paper [16] introduced the LASSO approach to linear models. Various properties of the method were demonstrated such as in [23] and extensions of the model to different frameworks were discussed in [11,12,24] etc.

The approach can be extended to the multivariate Bernoulli distribution since it is a member of the exponential family. What we have to do is to apply the l_1 penalty to the coefficients in (5.1), and the target function is

$$L_\lambda(x,y) = \frac{1}{n}\sum_{i=1}^{n} l\big(y(i), \mathbf{f}\big(x(i)\big)\big) + \sum_{\tau \in \mathcal{T}} \lambda_\tau \sum_{j=1}^{p} \big|c_j^\tau\big|, \tag{5.6}$$

where λ_τ are the tuning parameters need to be chosen adaptively. The superscript τ allows flexibility to have natural parameters with different levels of complexity. For tuning in penalized regression problems, the randomized generalized approximate cross-validation (GACV) designed for smoothing spline models introduced in [21] can be derived for LASSO problem, such as in [15]. The widely used information criterion AIC and BIC can also be implemented, but the degrees of freedom cannot be calculated exactly. [9] demonstrates that the number of nonzero estimates can serve as a good approximation in the multivariate Bernoulli logistic model. There are several efficient algorithms proposed to optimize the problem (5.6), for example, the LASSO-pattern search introduced in [15] can handle large number of unknowns provided that it is known that at most a modest number are nonzeros. Recently, [14] has extended the algorithm in [15] to the scale of multi-millions of unknowns. Coordinate descent [5] is also proven to be fast in solving large p small n problems.

5.5. Smoothing spline ANOVA model

The smoothing spline model gained popularity in non-linear statistical inference since it was proposed in [2] for univariate predictor variables. More importantly, multiple smoothing spline models for generalized linear models enable researchers to study complex real world data sets with increasingly powerful computers as described in [18].

As a member of the exponential family, the multivariate Bernoulli distribution can be formulated under smoothing spline ANOVA framework. [6] considers the smoothing spline ANOVA multivariate Bernoulli model but the interactions are restricted to be constant. However, in general the natural parameters or linear predictors f's can be relaxed to reside in a reproducing kernel Hilbert space. That is to say, for the observed predictor vector x, we have

$$f^\tau(x) = \eta^\tau(x) \qquad \text{with } \eta^\tau \in \mathcal{H}^\tau, \tau \in \mathcal{T}, \tag{5.7}$$

where $\mathcal{H}^\tau$ is a reproducing kernel Hilbert space and the superscript τ allows a more flexible model such that the natural parameters can come from different reproducing kernel Hilbert spaces. Further, $\mathcal{H}^\tau$ can be formulated to have several components to handle multivariate predictor variables, that is $\mathcal{H}^\tau = \oplus_{\beta=0}^{p} \mathcal{H}_\beta^\tau$ and details can be found in [7].

As a result, the η^τ is estimated from the variational problem

$$\mathcal{I}_\lambda(x, y) = \frac{1}{n} \sum_{i=1}^{n} l\big(y(i), \boldsymbol{\eta}\big(x(i)\big)\big) + \lambda J(\boldsymbol{\eta}), \tag{5.8}$$

where $\boldsymbol{\eta}$ is the vector form of η^τ's. The penalty is seen to be

$$\lambda J(\boldsymbol{\eta}) = \lambda \sum_{\tau \in \mathcal{T}} \theta_\tau^{-1} \| P_1^\tau \eta^\tau \|^2 \tag{5.9}$$

with λ and θ_τ being the smoothing parameters. This is an over-parameterization adopted in [7], as what really matters are the ratios λ/θ_τ. The functional P_1^τ projects function η^τ in $\mathcal{H}^\tau$ onto the smoothing subspace $\mathcal{H}_1^\tau$.

By the argument of smoothing spline ANOVA model in [7], the minimizer η^τ has the expression as in [17],

$$\eta^\tau(x) = \sum_{\nu=1}^{m} d_\nu^\tau \phi_\nu^\tau(x) + \sum_{i=1}^{n} c_i^\tau R^\tau(x_i, x), \tag{5.10}$$

where $\{\phi_\nu^\tau\}_{\nu=1}^m$ is a basis of $\mathcal{H}_0^\tau = \mathcal{H}^\tau \ominus \mathcal{H}_1^\tau$, the null space corresponding to the projection functional P_1^τ. $R^\tau(\cdot, \cdot)$ is the reproducing kernel for $\mathcal{H}_1^\tau$.

The variational problem (5.8) utilizing the smoothing spline ANOVA framework can be solved by iterative re-weighted least squares (5.5) due to the linear formulation (5.10). More on tuning and computations including software will appear in [3].

6. Conclusion

We have shown that the multivariate Bernoulli distribution, as a member of the exponential family, is a way to formulate the graph structure of binary variables. It can not only model the main effects and pairwise interactions as the Ising model does, but also is capable of estimating higher order interactions. Importantly, the independence structure of the graph can be modeled via significance of the natural parameters. The most interesting observation of the multivariate Bernoulli distribution is its similarity to the multivariate Gaussian distribution. Both of them have the property that independence and uncorrelatedness of the random variables are equivalent, which is generally not true for other distributions. In addition, the marginal and conditional distributions of a subset of variables still follow the multivariate Bernoulli distribution.

Furthermore, the multivariate Bernoulli logistic model extends the distribution to a generalized linear model framework to include effects of predictor variables. Under this model, the traditional statistical inferences such as point estimation, hypothesis test and confidence intervals can be implemented as discussed in [10].

Finally, we consider two extensions to the multivariate Bernoulli logistic model. First, the variable selection technique using LASSO can be incorporated to enable finding important patterns from a large number of candidate covariates. Secondly, the smoothing spline ANOVA model is introduced to consider non-linear effects of the predictor variables in nodes, edges and cliques level.

Appendix: Proofs

Proof of Proposition 2.1. With the joint density function of the random vector (Y_1, Y_2), the marginal distribution of Y_1 can be derived

$$\begin{aligned} P(Y_1 = 1) &= P(Y_1 = 1, Y_2 = 0) + P(Y_1 = 1, Y_2 = 1) \\ &= p_{10} + p_{11}. \end{aligned}$$

Similarly,

$$P(Y_1 = 0) = p_{00} + p_{11}.$$

Combining the side condition of the parameters p's,

$$P(Y_1 = 1) + P(Y_1 = 0) = p_{00} + p_{01} + p_{10} + p_{11} = 1.$$

This demonstrates that Y_1 follows the univariate Bernoulli distribution and its density function is (2.1).

Regarding the conditional distribution, notice that

$$\begin{aligned} P(Y_1 = 0 | Y_2 = 0) &= \frac{P(Y_1 = 0, Y_2 = 0)}{P(Y_2 = 0)} \\ &= \frac{p_{00}}{p_{00} + p_{10}}, \end{aligned}$$

and the same process can be repeated to get

$$P(Y_1 = 1 | Y_2 = 0) = \frac{p_{10}}{p_{00} + p_{10}}.$$

Hence, it is clear that with condition $Y_2 = 0$, Y_1 follows a univariate Bernoulli distribution as well. The same scenario can be examined for the condition $Y_2 = 1$. Thus, the conditional distribution of Y_1 given Y_2 is given as (2.11). □

Proof of Lemma 2.1. Expand the log-linear formulation of the bivariate Bernoulli distribution (2.9) into factors

$$P(Y_1 = y_1, Y_2 = y_2) = p_{00} \exp(y_1 f^1) \exp(y_2 f^2) \exp(y_1 y_2 f^{12}). \tag{A.1}$$

It is not hard to see that when $f^{12} = 0$, the density function (A.1) is separable to two components with only y_1 and y_2 in them. Therefore, the two random variables corresponding to the formula are independent. Conversely, when Y_1 and Y_2 are independent, their density function should be separable in terms of y_1 and y_2, which implies $y_1 y_2 f^{12} = 0$ for any possible values of y_1 and y_2. The assertion dictates that f^{12} is zero. □

Proof of Lemma 3.1. Consider the log-linear formulation (3.4), the natural parameters f's are combined with products of some components of y. Let us match terms in the $f^{j_1 \cdots j_r} B^{j_1 \cdots j_r}(y)$ from log-linear formulation (3.4) with the coefficient for the corresponding product $y_{j_1} \cdots y_{j_r}$ terms in (3.1). The exponents of p's in (3.1) can be expanded to summations of different products $B^\tau(y)$ with $\tau \in \mathcal{T}$ and all the p's with $y_{j_1}, \dots, y_{j_r}$ in the exponent have effect on $f^{j_1 \cdots j_r}$ so all the positions other than $j_1, \dots, j_r$ must be zero. Furthermore, those p's with positive $y_{j_1} \cdots y_{j_r}$ in its exponent appear in the numerator of $\exp[f^{j_1 \cdots j_r}]$ and the product is positive only if there are even number of 0's in the positions $j_1, \dots, j_r$. The same scenario applies to the p's with negative products in the exponents.

What's more, notice that $p_{00 \cdots 0} = b(\mathbf{f})$ and

$$\begin{aligned} \exp[S^{j_1 \cdots j_r}] &= \exp\Biggl[\sum_{1 \le s \le r} f^{j_s} + \sum_{1 \le s < t \le r} f^{j_s j_t} + \cdots + f^{j_1 j_2 \cdots j_r}\Biggr] \\ &= \prod_{1 \le s \le r} \exp[f^{j_s}] \prod_{1 \le s < t \le r} \exp[f^{j_s j_t}] \cdots \exp[f^{j_1 j_2 \cdots j_r}] \end{aligned} \tag{A.2}$$

and apply the formula for $\exp[f^{j_1 \cdots j_r}]$ with cancellation of terms in the numerators and the denominators. The resulting (3.6) can then be verified.

Finally, (3.7) is a trivial extension of (3.6) by exchanging the numerator and the denominator. □

Proof of Theorem 3.1. Here, we take use of the moment generating function (3.13) but it is also possible to directly work on the probability density function (3.1). The mgf can be rewritten as

$$\psi(\mu_1, \dots, \mu_K) = \frac{1}{\exp[b(\mathbf{f})]} \sum_{r=1}^{K} \sum_{j_1 \le j_2 \le \cdots \le j_r} \exp[S^{j_1 j_2 \cdots j_r}] \prod_{k=1}^{r} \exp[\mu_{j_k}]. \tag{A.3}$$

It is not hard to see that this is a polynomial function of the unknown variables $\exp(\mu_k)$ for $k = 1, \dots, K$. The independence of the random variables $Y_1, Y_2, \dots, Y_K$ is equivalent to that (A.3) can be separated into components of μ_k or equivalently $\exp(\mu_k)$.

(⇒) If the random vector Y is independent, the moment generating function should be separable and assume the formulation is

$$\psi(\mu_1, \ldots, \mu_K) = C \prod_{k=1}^{K} \bigl(\alpha_k + \beta_k \exp[\mu_k]\bigr), \tag{A.4}$$

where α_k and β_k are functions of parameters S's and C is a constant. If we expand (A.4) to polynomial function of $\exp[\mu_k]$ and determine the corresponding coefficients, (3.8) and (3.9) will be derived.

(⇐) Suppose (3.9) holds, then we have

$$\exp\bigl[S^{j_1 j_2 \cdots j_r}\bigr] = \prod_{k=1}^{r} \exp\bigl[f^{j_k}\bigr],$$

and as a result, the moment generating function can be decomposed to a product of components of $\exp[\mu_k]$ like (A.4) with the following relations

$$\begin{aligned} C &= \frac{1}{\exp[b(\mathbf{f})]}, \\ \alpha_k &= 1, \\ \beta_k &= \exp\bigl[f^k\bigr]. \end{aligned}$$

□

Proof of Theorem 3.2. The idea of proving the group independence of multivariate Bernoulli variables are similar to Theorem 3.1. Instead of decomposing the moment generating function to products of μ_k, we only have to separate them into groups with each only involving the dependent random variables. That is to say, the moment generating function with two separately independent nodes in the multivariate Bernoulli should have the form

$$\begin{aligned} &\psi(\mu_1, \ldots, \mu_K) \\ &\quad = \bigl(\alpha_0 + \alpha_1 \exp[\mu_1] + \cdots + \alpha_r \exp[\mu_r]\bigr) \cdot \bigl(\beta_0 + \beta_1 \exp[\mu_{r+1}] + \cdots + \beta_s \exp[\mu_K]\bigr). \end{aligned}$$

Matching the corresponding coefficients of this separable moment generating function and the natural parameters leads to the conclusion (3.10). □

Acknowledgements

Research of all three authors was supported in part by NIH Grant EY09946 and NSF Grant DMS-09-06818.

References

[1] Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9** 485–516. MR2417243

[2] Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31** 377–403. MR0516581

[3] Dai, B. (2012). Multivariate Bernoulli distribution models. Technical report. Dept. Statistics, Univ. Wisconsin, Madison, WI 53706.

[4] Ding, S., Wahba, G. and Zhu, X. (2011). Learning higher-order graph structure with features by structure penalty. In *Advances in Neural Information Processing Systems* **24** 253–261. 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12–14 December 2011, Granada, Spain.

[5] Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **44** 1–22.

[6] Gao, F., Wahba, G., Klein, R. and Klein, B. (2001). Smoothing spline ANOVA for multivariate Bernoulli observations, with application to ophthalmology data. *J. Amer. Statist. Assoc.* **96** 127–160. MR1952725

[7] Gu, C. (2002). *Smoothing Spline ANOVA Models*. *Springer Series in Statistics*. New York: Springer. MR1876599

[8] Ising, E. (1925). Beitrag zur Theorie des Ferromagnetismus. *Z. Phys.* **31** 253–258.

[9] Ma, X. (2010). Penalized regression in reproducing kernel Hilbert spaces with randomized covariate data. Technical Report No. 1159. Dept. Statistics, Univ. Wisconsin, Madison, WI 53706.

[10] McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. New York: Chapman & Hall.

[11] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[12] Park, T. and Casella, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103** 681–686. MR2524001

[13] Ravikumar, P., Wainwright, M.J. and Lafferty, J.D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319. MR2662343

[14] Shi, W., Wahba, G., Irizarry, R., Corrado Bravo, H. and Wright, S. (2012). The partitioned LASSO-patternsearch algorithm with application to gene expression data. *BMC Bioinformatics* **13** 98–110.

[15] Shi, W., Wahba, G., Wright, S., Lee, K., Klein, R. and Klein, B. (2008). LASSO-Patternsearch algorithm with application to ophthalmology and genomic data. *Stat. Interface* **1** 137–153. MR2425351

[16] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58** 267–288. MR1379242

[17] Wahba, G. (1990). *Spline Models for Observational Data*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **59**. Philadelphia, PA: SIAM. MR1045442

[18] Wahba, G., Wang, Y., Gu, C., Klein, R. and Klein, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *Ann. Statist.* **23** 1865–1895. MR1389856

[19] Wainwright, M. and Jordan, M. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* **1** 1–305.

[20] Whittaker, J. (1990). *Graphical Models in Applied Mathematical Multivariate Statistics*. New York: Wiley.

[21] Xiang, D. and Wahba, G. (1994). A generalized approximate cross validation for smoothing splines with non-Gaussian data. Technical Report No. 930. Dept. Statistics, Univ. Wisconsin, Madison, WI 53706.

[22] Xue, L., Zou, H. and Cai, T. (2012). Nonconcave penalized composite conditional likelihood estimation of sparse Ising models. *Ann*. *Statist*. **40** 1403–1429. MR3015030

[23] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J*. *Mach*. *Learn*. *Res*. **7** 2541–2563. MR2274449

[24] Zhao, P. and Yu, B. (2007). Stagewise lasso. *J*. *Mach*. *Learn*. *Res*. **8** 2701–2726. MR2383572