Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Multivariate Analysis 97 (2006) 1573 – 1585

Journal of Multivariate Analysis

www.elsevier.com/locate/jmva

A matrix-valued Bernoulli distribution

Gianfranco Lovison[1]

Dipartimento di Scienze Statistiche e Matematiche "S. Vianelli", Università di Palermo Viale delle Scienze 90128 Palermo, Italy

Received 17 September 2004
Available online 10 August 2005

Abstract

Matrix-valued distributions are used in continuous multivariate analysis to model sample data matrices of continuous measurements; their use seems to be neglected for binary, or more generally categorical, data. In this paper we propose a matrix-valued Bernoulli distribution, based on the log-linear representation introduced by Cox [The analysis of multivariate binary data, Appl. Statist. 21 (1972) 113–120] for the Multivariate Bernoulli distribution with correlated components.

AMS 1991 subject classification: 62E05; 62Hxx

Keywords: Correlated multivariate binary responses; Multivariate Bernoulli distribution; Matrix-valued distributions

1. Introduction

Matrix-valued distributions are used in continuous multivariate analysis (see, for example, [10]) to model sample data matrices of continuous measurements, allowing for both variable-dependence and unit-dependence. Their potentials seem to have been neglected for binary, and more generally categorical, data. This is somewhat surprising, since the natural, elementary representation of datasets with categorical variables is precisely in the form of sample binary data matrices, through the 0-1 coding of categories.

E-mail address: lovison@unipa.it
URL: http://dssm.unipa.it/lovison.

[1] Work supported by MIUR 60% 2000, 2001 and MIUR P.R.I.N. 2002 grants.

0047-259X/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2005.06.008

In this paper we denote an observed binary data matrix by $\boldsymbol{Z}$ and the matrix-valued random variable, of which $\boldsymbol{Z}$ is a sample realization, by $\mathcal{Z}$. Both are $n \times p$ matrices, where n is the number of sample units and p is the number of binary variables. As long as the units can be assumed to be independent, it is appropriate to model sufficient statistics obtained by marginalizing $\mathcal{Z}$, or its appropriate functions, over units; this kind of marginalization underlies the practice of directly modeling the p-dimensional frequency table obtained cross-classifying the p binary variables. If units are dependent, e.g. because they are sampled through a complex sampling scheme or come from a longitudinal or spatial study, then both the dependence structure of variables and that of units, and even the cross-dependence of variables and units, can be of interest. We shall call these three dimensions of dependence *pure unit-dependence*, *pure variable-dependence* and *mixed unit/variable-dependence*. When all these dimensions are actually of interest, it would seem natural to analyze $\boldsymbol{Z}$ by setting up a parametric model for $Pr(\mathcal{Z} = \boldsymbol{Z})$, which can take into account simultaneously all these three types of dependence, in much the way the Matrix Normal distribution does in the continuous case. Applications motivating the search for such a model can be found in many fields, for example in epidemiology, when family members are examined for the presence of multiple diseases (see, e.g., [3]), and in toxicology, when the offspring of treated pregnant animals are assessed for multiple outcomes (see, e.g., [9]).

The objective of this paper is to show the advantages of specifying a binary matrix distribution directly for $\mathcal{Z}$. Since the characterization of the Matrix Bernoulli distribution proposed in this paper parallels that used by many authors for the Matrix Normal, we shall devote the next section to a brief review of the Matrix Normal distribution. Section 3 presents the characterization of the Matrix Bernoulli distribution in general terms and Section 4 illustrates three special cases obtained by making simplifying assumptions about the dependence structure among units. In Section 5 some results in maximum likelihood estimation of the parameters of the Matrix Bernoulli distribution are described. Finally, Section 6 contains a discussion of possible extensions, and of the difficulties linked with them.

2. A brief review of the Matrix Normal distribution

The Matrix Normal is by far the most studied matrix-valued distribution; a thorough treatment can be found in [8]. Here we just give some basic results that will be useful in the sequel.

There are several ways to characterize the Matrix Normal distribution; the most useful one for our subsequent developments is given in the following definition (see, e.g., [8, p. 55]):

A random matrix $\boldsymbol{X}$ of continuous data is said to have a Matrix Normal distribution, denoted by $\boldsymbol{X} \sim \mathcal{N}_{n,p}(\boldsymbol{M}, \boldsymbol{\Xi}, \boldsymbol{\Sigma})$, if $vec(\boldsymbol{X})$ has a Multivariate Normal distribution with parameters $vec(\boldsymbol{M})$ and $\boldsymbol{\Xi} \otimes \boldsymbol{\Sigma}$, i.e. $vec(\boldsymbol{X}) \sim \mathcal{N}_{np}(vec(\boldsymbol{M}), \boldsymbol{\Xi} \otimes \boldsymbol{\Sigma})$.

By this approach, the Matrix Normal distribution of a random matrix $\boldsymbol{X}$ is derived from the Multivariate Normal distribution of its vectorized form. From this definition, it is straightforward to deduce that $\boldsymbol{X}$ has p.d.f.:

$$f(\boldsymbol{X}) = \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Xi}|^{p/2}|\boldsymbol{\Sigma}|^{n/2}} \exp\left\{-\frac{1}{2}tr[\boldsymbol{\Xi}^{-1}(\boldsymbol{X}-\boldsymbol{M})\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{M})^T]\right\}, \tag{1}$$

where $\boldsymbol{M}$, $\boldsymbol{\Xi}$ and $\boldsymbol{\Sigma}$ are $n \times p$, $n \times n$ and $p \times p$ matrices, respectively.

The three matrix-valued parameters $\boldsymbol{M}$, $\boldsymbol{\Xi}$ and $\boldsymbol{\Sigma}$ have intuitive interpretations. Clearly $\boldsymbol{M}$ is the (matrix-valued) expected value of $\boldsymbol{X}$; as for $\boldsymbol{\Xi}$ and $\boldsymbol{\Sigma}$, it is interesting to think of $\boldsymbol{\Xi}$ as representing the covariance between the rows of $\boldsymbol{X}$, and $\boldsymbol{\Sigma}$ as representing the covariance between the columns of $\boldsymbol{X}$. This interpretation is particularly useful when $\boldsymbol{X}$ is a sample data matrix of p-variate observations on n sampling units. Then $\boldsymbol{\Xi}$ can be thought of as containing the "pure unit-dependence" parameters and $\boldsymbol{\Sigma}$ the "pure variable-dependence" parameters. Moreover, the two aspects, unit-dependence and variable-dependence, do not interact, in the sense that

$$cov(X_{ij}, X_{hk}) = \xi_{ih}\sigma_{jk}. \tag{2}$$

The form of the covariance in (2) is a consequence of the absence of mixed unit/variable interaction implicit in (1): the relationship between two variables is the same on any unit, and the relationship between two units is the same on any variable. This is better seen in terms of correlations instead of covariances. Using $var(X_{ij}) = \xi_{ii}\sigma_{jj}$ and (2), we obtain:

$$\begin{aligned}\rho(X_{ij}, X_{ik}) &= \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}}\sqrt{\sigma_{kk}}} = \rho(X_{hj}, X_{hk}) \quad \forall i, h,\\ \rho(X_{ij}, X_{hj}) &= \frac{\xi_{ih}}{\sqrt{\xi_{ii}}\sqrt{\xi_{hh}}} = \rho(X_{ik}, X_{hk}) \quad \forall j, k.\end{aligned}$$

On the other hand, the multiplicative form of (2) also implies a lack of identifiability of (1): if $\boldsymbol{X} \sim \mathcal{N}_{n,p}(\boldsymbol{M}, \boldsymbol{\Xi}, \boldsymbol{\Sigma})$ then $\boldsymbol{X}$ is also distributed as $\mathcal{N}_{n,p}(\boldsymbol{M}, \frac{1}{a}\boldsymbol{\Xi}, a\boldsymbol{\Sigma})$ for any positive scalar a. So, in a sense, unit-dependence and variable-dependence are never completely distinguishable in their contribution to the overall dependence structure of $\boldsymbol{X}$.

3. The matrix-valued Bernoulli distribution

In this section we set out to characterize the matrix-valued Bernoulli distribution by analogy with the Matrix Normal. To do so, it is useful to begin by recalling some elementary results on the Multivariate Bernoulli distribution.

3.1. The multivariate Bernoulli distribution

Let Z^{A_k} be a binary response, which measures whether a dichotomous variable A_k is present ('success') or absent ('failure'):

$$Z^{A_k} = \begin{cases} 1 & \text{if a success is recorded on variable } A_k,\ k = 1, \ldots, m,\\ 0 & \text{otherwise.}\end{cases}$$

Consider the random vector $\boldsymbol{z}^A = [Z^{A_1}, \ldots, Z^{A_k}, \ldots, Z^{A_m}]^T$ and its realization $\boldsymbol{z}^A = [z^{A_1}, \ldots, z^{A_k}, \ldots, z^{A_m}]^T$. There are many ways to model the possible dependence among components in $\boldsymbol{z}^A$ (see, for different approaches: [2,4,1,6,12,11]). The one which leads more easily to a matrix-valued generalization is that proposed by Cox [4], who introduced the following parameterization for the distribution of $\boldsymbol{z}$:

$$\begin{aligned} Pr(\boldsymbol{z}^A = z^A) = C \exp \Bigg[& \sum_{k}^{m} \psi^{A_k} z^{A_k} + \sum_{k \neq h}^{m} \psi^{A_k A_h} z^{A_k} z^{A_h} \\ & + \sum_{k \neq h \neq r}^{m} \psi^{A_k A_h A_r} z^{A_k} z^{A_h} z^{A_r} \\ & + \cdots + \psi^{A_1, A_2, \ldots, A_m} z^{A_1} z^{A_2} \ldots z^{A_m} \Bigg], \end{aligned} \tag{3}$$

where C is a normalizing constant.

Parameterization (3) is completely general. The actual meaning of the ψ parameters will depend on the nature of the binary variables Z^{A_k}'s. For example, focussing on the simple case of two variables: if Z^{A_k}, Z^{A_h} are two *different binary variables recorded on each sampling unit*, then $\psi^{A_k A_h}$ is a "pure variable-association" parameter; if Z^{A_k}, Z^{A_h} refer to two *sampling draws from the same binary variable*, then $\psi^{A_k A_h}$ is a "pure unit-association" parameter; and finally, if Z^{A_k} is a variable recorded on a sampling unit and Z^{A_h} is a different variable recorded on a different unit, then $\psi^{A_k A_h}$ is a "mixed variables/units-association" parameter. In what follows, we shall stress this difference by using different symbols: θ for the first type of parameters, λ for the second and ϕ for the third.

3.2. The general matrix-valued Bernoulli distribution: vector representation

To keep things simple, we shall illustrate the matrix-variate case in a relatively simple setting, i.e. with only two variables and n sampling units, i.e. with $p = 2$ and arbitrary n. In order to avoid double subscripts, we shall use the letters A and B, instead of A_1 and A_2, to denote the two variables.

Then, let Z^A, Z^B be two binary responses, which measure whether the two dichotomous variables of interest A and B are present ('success') or absent ('failure') for n sample units:

$$Z_i^A = \begin{cases} 1 & \text{if the } i\text{th unit is a success on variable } A, \ i = 1, \ldots, n, \\ 0 & \text{otherwise.} \end{cases}$$

$$Z_i^B = \begin{cases} 1 & \text{if the } i\text{th unit is a success on variable } B, \ i = 1, \ldots, n, \\ 0 & \text{otherwise} \end{cases}$$

and denote by $\boldsymbol{z}^A = [Z_1^A, \ldots, Z_i^A, \ldots, Z_n^A]^T$ and $\boldsymbol{z}^B = [Z_1^B, \ldots, Z_i^B, \ldots, Z_n^B]^T$ the two random vectors which generate the realizations $z^A = [z_1^A, \ldots, z_i^A, \ldots, z_n^A]^T$ and $z^B = [z_1^B, \ldots, z_i^B, \ldots, z_n^B]^T$.

The natural arrangement for the two vectors $\boldsymbol{z}^A$ and $\boldsymbol{z}^B$ would be as columns of the random matrix:

$$\boldsymbol{\mathcal{Z}} = [\boldsymbol{z}^A, \boldsymbol{z}^B] = \begin{bmatrix} Z_1^A & Z_1^B \\ Z_2^A & Z_2^B \\ \vdots & \vdots \\ Z_i^B & Z_i^B \\ \vdots & \vdots \\ Z_n^A & Z_n^B \end{bmatrix}.$$

However, in order to follow an approach similar to that used in the definition given in Section 2 for the Matrix Normal distribution, we first consider the distribution of the vectorized form of $\boldsymbol{\mathcal{Z}}^T$:

$$vec(\boldsymbol{\mathcal{Z}}^T) = \left[Z_1^A, Z_1^B, Z_2^A, Z_2^B, \ldots, Z_i^B, Z_i^B, \ldots, Z_n^A, Z_n^B\right]^T.$$

The (transposed) observed matrix $\mathbf{Z}^T$ is vectorized accordingly:

$$vec(\mathbf{Z}^T) = \left[z_1^A, z_1^B, z_2^A, z_2^B, \ldots, z_i^B, z_i^B, \ldots, z_n^A, z_n^B\right]^T.$$

We now need to accommodate the possible dependence among variables, units and variables/units. In order to do so, we extend (3) and introduce the following parameterization for the distribution of $vec(\boldsymbol{\mathcal{Z}}^T)$:

$$\begin{aligned} Pr(vec(\boldsymbol{\mathcal{Z}}^T) &= vec(\mathbf{Z}^T)) \\ &= C \exp\Bigg[\sum_i^n \theta_i^A z_i^A + \sum_i^n \theta_i^B z_i^B + \sum_i^n \theta_i^{AB} z_i^A z_i^B \\ &\quad + \sum_{i \neq h}^n \lambda_{ih}^{AA} z_i^A z_h^A + \sum_{i \neq h}^n \phi_{ih}^{AB} z_i^A z_h^B + \sum_{i \neq h}^n \lambda_{ih}^{BB} z_i^B z_h^B \\ &\quad + \sum_{i \neq h \neq r}^n \lambda_{ihr}^{AAA} z_i^A z_h^A z_r^A + \sum_{i \neq h \neq r}^n \phi_{ihr}^{AAB} z_i^A z_h^A z_r^B \\ &\quad + \cdots + \sum_{i \neq h \neq r}^n \phi_{ihr}^{BBA} z_i^B z_h^B z_r^A + \sum_{i \neq h \neq r}^n \lambda_{ihr}^{BBB} z_i^B z_h^B z_r^B \\ &\quad + \cdots + \lambda_{1,2,\ldots,n}^{AA\ldots A} z_1^A z_2^A \cdots z_n^A + \cdots + \lambda_{1,2,\ldots,n}^{BB\ldots B} z_1^B z_2^B \cdots z_n^B\Bigg], \end{aligned} \tag{4}$$

where C is again a normalizing constant.

As already mentioned, in (4) we can distinguish three types of parameters:

- θ_i^A, θ_i^B and θ_i^{AB} are usual log-linear parameters referring to the association structure of the variables, e.g.:

  $$\theta_i^A = \log\left\{\frac{Pr[Z_i^A = 1|rest = 0]}{Pr[Z_i^A = 0|rest = 0]}\right\},$$
  $$\theta_i^{AB} = \log\left\{\frac{Pr[Z_i^A = 1, Z_i^B = 1|rest = 0]Pr[Z_i^A = 0, Z_i^B = 0|rest = 0]}{Pr[Z_i^A = 0, Z_i^B = 1|rest = 0]Pr[Z_i^A = 1, Z_i^B = 0|rest = 0]}\right\}$$

  and therefore can be considered as "pure variable-association" parameters; in this respect, they play the same role as the μ_{ij} and σ_{jk} parameters in the Matrix Normal distribution. Although in theory these parameters might be subject-specific, they are usually assumed to be common to all units:

  $$\theta_i^A = \theta^A, \ \theta_i^B = \theta^B \ \text{and} \ \theta_i^{AB} = \theta^{AB} \quad \forall i.$$

- λ-parameters, i.e. parameters with one-variable repeated superscripts, like λ_{ih}^{AA}, λ_{ih}^{BB}, λ_{ihr}^{AAA}, λ_{ihr}^{BBB}, etc., refer to the intra-units dependence with respect to each variable, e.g.:

  $$\lambda_{ih}^{AA} = \log\left\{\frac{Pr[Z_i^A = 1, Z_h^A = 1|rest = 0]Pr[Z_i^A = 0, Z_h^A = 0|rest = 0]}{Pr[Z_i^A = 0, Z_h^A = 1|rest = 0]Pr[Z_i^A = 1, Z_h^A = 0|rest = 0]}\right\}$$

  and therefore can be considered as "pure unit-association" parameters; from this perspective, they are the analog of the ξ_{ih} parameters in the Matrix Normal distribution. These parameters satisfy symmetry constraints within each t-uple of units, for $2 \leqslant t \leqslant n$, but they can be different in different t-uples:

  $$\lambda_{ih}^{AA} = \lambda_{hi}^{AA} \quad \forall i, h,$$
  $$\lambda_{ihr}^{AAA} = \lambda_{irh}^{AAA} = \lambda_{hir}^{AAA} = \lambda_{hri}^{AAA} = \lambda_{rih}^{AAA} = \lambda_{rhi}^{AAA} \quad \forall i, h, r$$

  and so on.

- ϕ-parameters, i.e. parameters with two-variables superscripts, like ϕ_{ih}^{AB}, ϕ_{ihr}^{AAB}, ϕ_{ihr}^{BBA}, etc., refer to the intra-units dependence with respect to a specified combination of the two variables, e.g.:

  $$\phi_{ih}^{AB} = \log\left\{\frac{Pr[Z_i^A = 1, Z_h^B = 1|rest = 0]Pr[Z_i^A = 0, Z_h^B = 0|rest = 0]}{Pr[Z_i^A = 0, Z_h^B = 1|rest = 0]Pr[Z_i^A = 1, Z_h^B = 0|rest = 0]}\right\}.$$

  As such, they measure the "mixed variables/units-association" on a log-linear scale. Notice that by definition they do not have any analog in the Matrix Normal distribution, given the absence of variables/units interaction that is implicit in (1). These parameters also satisfy symmetry constraints with respect to simultaneous permutations of variables and units, e.g.:

  $$\phi_{ih}^{AB} = \phi_{hi}^{BA} \quad \forall i, h.$$

On the other hand, they need not satisfy symmetry constraints with respect to permutations of the variables alone or of the units alone, i.e. ϕ_{ih}^{AB} can be assumed to be equal or different from ϕ_{hi}^{AB} (and from ϕ_{ih}^{BA}), depending on the applications.

3.3. A matrix-valued Bernoulli distribution with quadratic exponential dependence structure

In general, it is not easy to write explicitly the p.d.f. of $\mathcal{Z}$ as a function of the sample matrix **Z** starting from (4), since the presence of second- and higher-order interactions among variables, among units and among variables/units implies the use of arrays with three or more dimensions. A special case which only involves the use of matrices is obtained by enforcing the condition that all interactions (among variables, among units, and among variables/units) of order greater than 1 equal 0. Of course, since in our illustration we have $p = 2$ variables, this condition puts no restrictions here as far as the variables are concerned, but it would constrain the type of variables-association admitted if we had $p > 2$. For units, and variables/units, this condition implies that only pairwise interactions are allowed, a severe constraint, that anyway encompasses many cases of applicative interest, as we shall see in Section 4.

The "only pairwise interactions" condition implies that the vectorized form of **Z** follows a quadratic exponential model (see [13,5]):

$$Pr(vec(\mathcal{Z}) = vec(\mathbf{Z})) = C(\mathbf{\Psi}) \exp\{vec(\mathbf{Z}^T)^T \mathbf{\Psi}\, vec(\mathbf{Z}^T)\}, \tag{5}$$

where $\mathbf{\Psi}$ is a partitioned matrix:

$$\mathbf{\Psi} = \begin{bmatrix} \mathbf{\Theta} & \mathbf{\Lambda}_{12} & \dots & \mathbf{\Lambda}_{1,n-1} & \mathbf{\Lambda}_{1,n} \\ \mathbf{\Lambda}_{21} & \mathbf{\Theta} & \dots & \mathbf{\Lambda}_{2,n-1} & \mathbf{\Lambda}_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{\Lambda}_{n-1,1} & \mathbf{\Lambda}_{n-1,2} & \dots & \mathbf{\Theta} & \mathbf{\Lambda}_{n-1,n} \\ \mathbf{\Lambda}_{n,1} & \mathbf{\Lambda}_{n,2} & \dots & \mathbf{\Lambda}_{n,n-1} & \mathbf{\Theta} \end{bmatrix} = \boldsymbol{I}_n \otimes \mathbf{\Theta} + \sum_{i}^{n} \sum_{j \neq i}^{n} \boldsymbol{E}_{ij} \otimes \mathbf{\Lambda}_{ij} \tag{6}$$

and $C(\mathbf{\Psi})$ is a normalizing constant. Since there are 2^{np} possible matrices in support of $\mathcal{Z}$, such a normalizing constant is given by

$$C(\mathbf{\Psi}) = \left[\sum_{k=1}^{2^{np}} \exp\{vec(\mathbf{Z}_k^T)^T \mathbf{\Psi}\, vec(\mathbf{Z}_k^T)\} \right]^{-1}.$$

In (6) $\boldsymbol{I}_n$ is the $n \times n$ identity matrix, $\boldsymbol{E}_{ij}$ is the (i, j) elementary matrix of order $n \times n$, and

$$\mathbf{\Theta} = \begin{bmatrix} \theta^A & \theta^{AB} \\ \theta^{AB} & \theta^B \end{bmatrix}, \quad \mathbf{\Lambda}_{ij} = \mathbf{\Lambda}_{ji}^T = \begin{bmatrix} \lambda_{ij}^A & \phi_{ij}^{AB} \\ \phi_{ij}^{BA} & \lambda_{ij}^B \end{bmatrix}.$$

By using the same characterization recalled in the definition of the Matrix Normal distribution, we can introduce the following definition:

The random matrix $\boldsymbol{Z}$ is said to have a Matrix Bernoulli distribution, denoted by $\boldsymbol{Z} \sim Ber_{n,p}(\boldsymbol{\Theta}, \boldsymbol{\Lambda}_{12}, \ldots, \boldsymbol{\Lambda}_{n-1,n})$, if $vec(\boldsymbol{Z})$ has a Multivariate Bernoulli distribution with parameter $\boldsymbol{\Psi}$, i.e. $vec(\boldsymbol{Z}) \sim Ber_{np}(\boldsymbol{\Psi})$.

Using a standard result on the properties of the Kronecker product and the trace function, we finally obtain from (5) an expression for $Pr(\mathcal{Z} = \boldsymbol{Z})$ as a function of $\boldsymbol{Z}$:

$$\begin{aligned} Pr(\mathcal{Z} = \boldsymbol{Z}) = {} & C(\boldsymbol{\Theta}, \boldsymbol{\Lambda}_{12}, \ldots, \boldsymbol{\Lambda}_{n-1,n}) \\ & \times \exp\left\{ tr[\boldsymbol{Z\Theta Z}^T] + \sum_i^n \sum_{j \neq i}^n tr[\boldsymbol{E}_{ij}\boldsymbol{Z\Lambda}_{ij}\boldsymbol{Z}^T] \right\}. \end{aligned} \tag{7}$$

In (7) the normalizing constant is given by

$$C(\boldsymbol{\Theta}, \boldsymbol{\Lambda}_{12}, \ldots, \boldsymbol{\Lambda}_{n-1,n}) = [S(\boldsymbol{\Theta}, \boldsymbol{\Lambda}_{12}, \ldots, \boldsymbol{\Lambda}_{n-1,n})]^{-1},$$

where

$$S(\boldsymbol{\Theta}, \boldsymbol{\Lambda}_{12}, \ldots, \boldsymbol{\Lambda}_{n-1,n}) = \sum_{k=1}^{2^{np}} \exp\left\{ tr[\boldsymbol{Z}_k\boldsymbol{\Theta Z}_k^T] + \sum_i^n \sum_{j \neq i}^n tr[\boldsymbol{E}_{ij}\boldsymbol{Z}_k\boldsymbol{\Lambda}_{ij}\boldsymbol{Z}_k^T] \right\}. \tag{8}$$

It is immediately seen, by simple inspection of (7), that the Matrix Bernoulli distribution does not suffer the lack of identifiability typical of the Matrix Normal distribution. On the other hand, unlike the Matrix Normal, the Matrix Bernoulli is not closed with respect to marginalization. Its use is therefore recommended when conditional, rather than marginal, associations between variables, units and variables/units are of substantive interest.

4. Special cases

It is useful to make some simplifying assumptions which lead to meaningful reductions of (7). In particular, in the applications it is usually possible to assume that the pure unit-association parameters (and the mixed variables/units interactions parameters, if present) take on some simplified form that reflects the association structure appropriate for the mechanism generating the data at hand. Three important examples are: (a) unit-independence, (b) unit-exchangeability, and (c) unit-Markovianity.

(a) *Unit-independence*: If the units are independent then $\boldsymbol{\Lambda}_{ij} = \boldsymbol{O}, \forall i, j$ and (7) reduce to

$$Pr(\mathcal{Z} = \boldsymbol{Z}) = C(\boldsymbol{\Theta}) \exp\{tr[\boldsymbol{Z\Theta Z}^T]\} = C(\boldsymbol{\Theta}) \exp\{tr[\boldsymbol{Z}^T\boldsymbol{Z\Theta}]\}. \tag{9}$$

We write $\boldsymbol{Z} \sim Ber_{n,p}(\boldsymbol{\Theta}, \boldsymbol{O})$ in short and say that $\boldsymbol{Z}$ is distributed as a *Standard Matrix Bernoulli*. It is instructive to re-write (9) in a more standard way by working out the normalizing constant. Let $y^A = \sum_i^n z_i^A = \sum_i^n (z_i^A)^2$ and $y^B = \sum_i^n z_i^B = \sum_i^n (z_i^B)^2$ be the sample marginal frequencies of successes on A and B, and $y^{AB} = \sum_i^n z_i^A z_i^B$ be the sample joint frequency of successes on A and B. We can write:

$$\mathbf{Z}^T\mathbf{Z} = \begin{bmatrix} \sum_i^n z_i^A & \sum_i^n z_i^A z_i^B \\ \sum_i^n z_i^A z_i^B & \sum_i^n z_i^B \end{bmatrix} = \begin{bmatrix} y^A & y^{AB} \\ y^{AB} & y^B \end{bmatrix}.$$

Recalling that $tr[\mathbf{AB}] = \sum_j \sum_k a_{kj} b_{jk}$, we get

$$\begin{aligned} C(\mathbf{\Theta}) \exp\{tr[\mathbf{Z}^T \mathbf{Z\Theta}]\} &= C(\mathbf{\Theta}) \exp\{\theta^A y^A + \theta^B y^B + 2\theta^{AB} y^{AB}\} \\ &= C(\mathbf{\Theta})[\exp(\theta^A)]^{(y^A - y^{AB})}[\exp(\theta^B)]^{(y^B - y^{AB})} \\ &\quad \times[\exp(\theta^A + \theta^B + 2\theta^{AB})]^{y^{AB}}, \end{aligned} \tag{10}$$

whence

$$C(\mathbf{\Theta}) = \frac{1}{[1 + \exp(\theta^A) + \exp(\theta^B) + \exp(\theta^A + \theta^B + 2\theta^{AB})]^n}. \tag{11}$$

Substituting (11) into (10) we can finally write:

$$Pr(\mathcal{Z} = \mathbf{Z}) = \pi_{00}^{n-y^A-y^B+y^{AB}} \pi_{01}^{y^B-y^{AB}} \pi_{10}^{y^A-y^{AB}} \pi_{11}^{y^{AB}}, \tag{12}$$

where

$$\pi_{00} = Pr(Z^A = 0, Z^B = 0) = \frac{1}{1 + \exp(\theta^A) + \exp(\theta^B) + \exp(\theta^A + \theta^B + \theta^{AB})},$$

$$\pi_{01} = Pr(Z^A = 0, Z^B = 1) = \frac{\exp(\theta^B)}{1 + \exp(\theta^A) + \exp(\theta^B) + \exp(\theta^A + \theta^B + \theta^{AB})},$$

$$\pi_{10} = Pr(Z^A = 1, Z^B = 0) = \frac{\exp(\theta^A)}{1 + \exp(\theta^A) + \exp(\theta^B) + \exp(\theta^A + \theta^B + \theta^{AB})},$$

$$\pi_{11} = Pr(Z^A = 1, Z^B = 1) = \frac{\exp(\theta^A + \theta^B + \theta^{AB})}{1 + \exp(\theta^A) + \exp(\theta^B) + \exp(\theta^A + \theta^B + \theta^{AB})}.$$

Clearly (12) is the p.d.f. of a Bivariate Bernoulli, with associated variables, on a sample of n i.i.d. units; more generally, in analogy with the Normal case, it is easy to show that modeling a binary matrix as a Standard Matrix Bernoulli is equivalent to considering its rows as n independent realizations from a Multivariate Bernoulli, as long as only first-order interactions are present among the variables.

(b) *Unit-exchangeability*: Suppose the n units have no serial order, but belong to the same cluster or matched set. It is then reasonable to assume not only that all unit-interactions and variable/unit-interactions of order higher than 1 are zero but also that the units are *exchangeable*, i.e. they have the same first-order unit-interaction parameters: $\lambda_{ih}^{AA} = \lambda^A, \lambda_{ih}^{BB} = \lambda^B \; \forall i, h$, and the same first-order variable/unit-interaction: $\phi_{ih}^{AB} = \phi^{AB} \; \forall i, h$. As a consequence: $\mathbf{\Lambda}_{ij} = \mathbf{\Lambda} \; \forall i \neq j$ with

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda^A & \phi^{AB} \\ \phi^{AB} & \lambda^B \end{bmatrix}.$$

Thus

$$\boldsymbol{\Psi} = \boldsymbol{I}_n \otimes \boldsymbol{\Theta} + (\boldsymbol{J}_n - \boldsymbol{I}_n) \otimes \boldsymbol{\Lambda}, \tag{13}$$

where $\boldsymbol{J}_n = \mathbf{1}_n \mathbf{1}_n^T$ is a square matrix of ones of order n, and

$$\begin{aligned} Pr(\mathcal{Z} = \mathbf{Z}) &= C(\boldsymbol{\Theta}, \boldsymbol{\Lambda}) \exp\{tr[\mathbf{Z}\boldsymbol{\Theta}\mathbf{Z}^T] + tr[(\boldsymbol{J}_n - \boldsymbol{I}_n)\mathbf{Z}\boldsymbol{\Lambda}\mathbf{Z}^T]\} \\ &= C(\boldsymbol{\Theta}, \boldsymbol{\Lambda}) \exp\{tr[\mathbf{Z}^T\mathbf{Z}\boldsymbol{\Theta}] + tr[\mathbf{Z}^T(\boldsymbol{J}_n - \boldsymbol{I}_n)\mathbf{Z}\boldsymbol{\Lambda}]\}. \end{aligned} \tag{14}$$

(c) *Unit-Markovianity*: Suppose now the n sample units are the outcome of a longitudinal study. Their dependence-structure is then serial in nature, and a simple way to take this structure into account is to assume it to be Markovian, i.e. that the unit-interaction and the mixed unit/variable-interaction parameters only depend on the distance of the unit labels: $\lambda_{ih}^{AA} = \lambda_{|h-i|}^{A}, \lambda_{ih}^{BB} = \lambda_{|h-i|}^{B}, \phi_{ih}^{AB} = \phi_{|h-i|}^{AB}\ \forall i, h$. In particular, suppose that first-order Markovianity holds, i.e. $\lambda_{|h-i|}^{A} = 0, \lambda_{|h-i|}^{B} = 0, \phi_{|h-i|}^{AB} = 0 \quad \forall\, |h-i| > 1$. This allows to write:

$$\boldsymbol{\Lambda}_{i,i+1} = \boldsymbol{\Lambda}_1 = \begin{bmatrix} \lambda_1^A & \phi_1^{AB} \\ \phi_1^{AB} & \lambda_1^B \end{bmatrix}, \ i = 1, \ldots, n-1, \quad \boldsymbol{\Lambda}_{i,i+t} = \boldsymbol{O}\ \forall t > 1.$$

Hence

$$\boldsymbol{\Psi} = \boldsymbol{I}_n \otimes \boldsymbol{\Theta} + \boldsymbol{L}_1 \otimes \boldsymbol{\Lambda}_1, \tag{15}$$

where $\boldsymbol{L}_1$ is the lag-one matrix of order $n \times n$:

$$\boldsymbol{L}_1 = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & 1 \\ 0 & 0 & 0 & \ldots & 0 & 0 \end{bmatrix}.$$

On using (15), the general form (7) reduces to

$$Pr(\mathcal{Z} = \mathbf{Z}) = C(\boldsymbol{\Theta}, \boldsymbol{\Lambda}_1) \exp\{tr[\mathbf{Z}^T\mathbf{Z}\boldsymbol{\Theta}] + tr[\mathbf{Z}^T\boldsymbol{L}_1\mathbf{Z}\boldsymbol{\Lambda}_1]\}. \tag{16}$$

5. Likelihood inference

One of the advantages of considering a matrix-valued distribution is the compactness it provides, and the possibility of employing matrix differentiation techniques in likelihood-based computations.

To begin with, the p.d.f. (7) immediately highlights the (jointly) sufficient statistics for the parameters. By writing it as

$$\begin{aligned} Pr(\mathcal{Z} = \mathbf{Z}) &= [S(\boldsymbol{\Theta}, \boldsymbol{\Lambda}_{12}, \ldots, \boldsymbol{\Lambda}_{n-1,n})]^{-1} \\ &\quad \times \exp\{tr[\mathbf{Z}^T\mathbf{Z}\boldsymbol{\Theta}]\} \prod_{i}^{n} \prod_{j \neq i}^{n} \exp\{tr[\mathbf{Z}^T\boldsymbol{E}_{ij}\mathbf{Z}\boldsymbol{\Lambda}_{ij}]\}, \end{aligned} \tag{17}$$

by the factorization criterion it is clear that $\boldsymbol{Z}^T\boldsymbol{Z}$ and $\boldsymbol{Z}^T\boldsymbol{E}_{ij}\boldsymbol{Z}, i, j = 1, \ldots, n, i \neq j$ are jointly sufficient statistics for the matrix-valued parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\Lambda}_{ij}, i, j = 1, \ldots, n, i \neq j$. Then, in particular:

(a) in the unit-independence model, the sufficient statistic for $\boldsymbol{\Theta}$ is $\boldsymbol{Z}^T\boldsymbol{Z}$;
(b) in the unit-exchangeability model, the jointly sufficient statistics for $\boldsymbol{\Theta}$ and $\boldsymbol{\Lambda}$ are $\boldsymbol{Z}^T\boldsymbol{Z}$ and $\boldsymbol{Z}^T(\boldsymbol{J}_n - \boldsymbol{I}_n)\boldsymbol{Z}$, respectively; and
(c) in the first-order Markovian model for the units, the jointly sufficient statistics for $\boldsymbol{\Theta}$ and $\boldsymbol{\Lambda}_1$ are $\boldsymbol{Z}^T\boldsymbol{Z}$ and $\boldsymbol{Z}^T\boldsymbol{L}_1\boldsymbol{Z}$, respectively.

Through the use of matrix derivative techniques, form (7) is also suitable for the derivation of results concerning maximum likelihood estimation. From (17), the log-likelihood for the parameters $\boldsymbol{\Theta}$, $\boldsymbol{\Lambda}_{ij}, i, j = 1, \ldots, n, i \neq j$ is

$$\ell(\boldsymbol{\Theta}, \boldsymbol{\Lambda}_{12}, \ldots, \boldsymbol{\Lambda}_{n-1,n}|\boldsymbol{Z}) = -\log\{S(\boldsymbol{\Theta}, \boldsymbol{\Lambda}_{12}, \ldots, \boldsymbol{\Lambda}_{n-1,n}) + tr[\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{\Theta}] + \sum_{i}^{n}\sum_{j\neq i}^{n} tr[\boldsymbol{Z}^T\boldsymbol{E}_{ij}\boldsymbol{Z}\boldsymbol{\Lambda}_{ij}]. \tag{18}$$

Let us denote by $s_k = \exp\left\{tr[\boldsymbol{Z}_k^T\boldsymbol{Z}_k\boldsymbol{\Theta}] + \sum_i^n \sum_{j\neq i}^n tr[\boldsymbol{Z}_k^T\boldsymbol{E}_{ij}\boldsymbol{Z}_k\boldsymbol{\Lambda}_{ij}]\right\}$ the kth generic term of the sum $S(\boldsymbol{\Theta}, \boldsymbol{\Lambda}_{12}, \ldots, \boldsymbol{\Lambda}_{n-1,n})$ defined in (8) and by $\hat{s}_k = \exp\left\{tr[\boldsymbol{Z}_k^T\boldsymbol{Z}_k\hat{\boldsymbol{\Theta}}] + \sum_i^n \sum_{j\neq i}^n tr[\boldsymbol{Z}_k^T\boldsymbol{E}_{ij}\boldsymbol{Z}_k\hat{\boldsymbol{\Lambda}}_{ij}]\right\}$ its maximum likelihood estimator. Using the rules of matrix differentiation (see [7]):

$$\frac{\partial\ell(\boldsymbol{\Theta}, \boldsymbol{\Lambda}_{12}, \ldots, \boldsymbol{\Lambda}_{n-1,n}|\boldsymbol{Z})}{\partial\boldsymbol{\Theta}} = -\frac{\sum_{k=1}^{2^{np}} s_k\boldsymbol{Z}_k^T\boldsymbol{Z}_k}{\sum_{k=1}^{2^{np}} s_k} + \boldsymbol{Z}^T\boldsymbol{Z}$$

and, for $i, j = 1, \ldots, n, \;\; j \neq i$:

$$\frac{\partial\ell(\boldsymbol{\Theta}, \boldsymbol{\Lambda}_{12}, \ldots, \boldsymbol{\Lambda}_{n-1,n}|\boldsymbol{Z})}{\partial\boldsymbol{\Lambda}_{ij}} = -\frac{\sum_{k=1}^{2^{np}} s_k\boldsymbol{Z}_k^T\boldsymbol{E}_{ij}\boldsymbol{Z}_k}{\sum_{k=1}^{2^{np}} s_k} + \boldsymbol{Z}^T\boldsymbol{E}_{ij}\boldsymbol{Z}.$$

Hence, the likelihood equations are

$$\frac{\sum_{k=1}^{2^{np}} \hat{s}_k\boldsymbol{Z}_k^T\boldsymbol{Z}_k}{\sum_{k=1}^{2^{np}} \hat{s}_k} = \boldsymbol{Z}^T\boldsymbol{Z}$$

and, for $i, j = 1, \ldots, n, \;\; j \neq i$:

$$\frac{\sum_{k=1}^{2^{np}} \hat{s}_k\boldsymbol{Z}_k^T\boldsymbol{E}_{ij}\boldsymbol{Z}_k}{\sum_{k=1}^{2^{np}} \hat{s}_k} = \boldsymbol{Z}^T\boldsymbol{E}_{ij}\boldsymbol{Z}.$$

But

$$\frac{\hat{s}_k}{\sum_{k=1}^{2^{np}} \hat{s}_k} = Pr(\boldsymbol{\mathcal{Z}} = \boldsymbol{Z}_k|\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Lambda}}_{12}, \ldots, \hat{\boldsymbol{\Lambda}}_{n-1,n}),$$

thus, the likelihood equations can be written as

$$E\left[\boldsymbol{Z}^T\boldsymbol{Z}|\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Lambda}}_{12}, \ldots, \hat{\boldsymbol{\Lambda}}_{n-1,n}\right] = \boldsymbol{Z}^T\boldsymbol{Z} \tag{19}$$

and, for $i, j = 1, \ldots, n, \ j \neq i$:

$$E\left[\boldsymbol{Z}^T\boldsymbol{E}_{ij}\boldsymbol{Z}|\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Lambda}}_{12}, \ldots, \hat{\boldsymbol{\Lambda}}_{n-1,n}\right] = \boldsymbol{Z}^T\boldsymbol{E}_{ij}\boldsymbol{Z}. \tag{20}$$

The likelihood equations in forms (19) and (20) provide an extension to the matrix-valued case of the well-known result by which the maximum likelihood estimates are the (unique) values that equate expected and observed sufficient statistics for the model.

6. Extensions

The main limitations of the Matrix Bernoulli distribution proposed in this paper are linked to its quadratic exponential dependence structure. As long as the "only pairwise interactions" condition holds, the extension of the results of the previous Sections is straightforward to situations such as:

- more than two binary variables;
- polytomous rather than dichotomous responses; and
- units grouped into clusters, with within-clusters unit-dependence but between-clusters independence.

As mentioned in Section 3.3, the extension to situations in which variables, units or variable/units dependence comprises higher-order interactions is conceptually easy but formally and computationally cumbersome, due to the need to work with higher dimensional arrays.

Finally, parameterizations of dependence alternative to the one introduced by Cox [4] and extended in this paper are going to be not only computationally but also conceptually challenging, since it is not immediately clear how they can be expressed in a form that leads to a matrix-valued distribution.

Acknowledgments

Part of this work has been done during a visit to the Statistical Laboratory of the University of Cambridge (UK). The author wishes to thank Pat Altham for the invitation that made this visit possible.

References

[1] P.M.E. Altham, Discrete variable analysis for individuals grouped into families, Biometrika 63 (1976) 263–269.

[2] R.R. Bahadur, A representation of the joint distribution of responses to *n* dichotomous items, in: H. Salomon (Ed.), Studies on Item Analysis and Prediction, Stanford University Press, Stanford CA, 1961, pp. 158–176.

[3] R.A. Betensky, A.S. Whittemore, An analysis of correlated multivariate binary data: application to familial cancers of the ovary and breast, Appl. Statist. 45 (1996) 411–429.

[4] D.R. Cox, The analysis of multivariate binary data, Appl. Statist. 21 (1972) 113–120.

[5] D.R. Cox, N. Wermuth, A note on the quadratic exponential binary distribution, Biometrika 81 (1994) 403–408.

[6] J.R. Dale, Global cross-ratio models for bivariate, discrete, ordered responses, Biometrics 42 (1986) 909–917.

[7] A. Graham, Kronecker Products and Matrix Calculus with Applications, Ellis Horwood, Chichester, 1981.

[8] A.K. Gupta, D.K. Nagar, Matrix Variate Distributions, Chapman & Hall/CRC, Boca Raton, 2000.

[9] M. Lefkopoulo, D. Moore, L. Ryan, The analysis of multiple correlated binary outcomes: application to rodent teratology experiments, J. Amer. Statist. Assoc. 84 (1989) 810–815.

[10] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, Academic Press, London, 1979.

[11] P.X.-K. Song, Multivariate dispersion models generated from Gaussian copula, Scand. J. Statist. 27 (2000) 305–320.

[12] J.L. Teugels, Some representations of the multivariate Bernoulli and binomial distributions, J. Multivariate Anal. 32 (1990) 256–268.

[13] L.P. Zhao, R.L. Prentice, Correlated binary regression using a quadratic exponential model, Biometrika 77 (1990) 642–648.