Correlated binomial models and correlation structures

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **39** (2006) 15365–15378 doi:10.1088/0305-4470/39/50/005

Correlated binomial models and correlation structures

Masato Hisakado[1], **Kenji Kitsukawa**[2] **and Shintaro Mori**[3]

[1] Standard and Poor's, Marunouchi 1-6-5, Chiyoda-ku, Tokyo 100-0005, Japan
[2] Graduate School of Media and Governance, Keio University, Endo 5322, Fujisawa, Kanagawa 252-8520, Japan
[3] Department of Physics, School of Science, Kitasato University, Kitasato 1-15-1, Sagamihara, Kanagawa 228-8555, Japan

E-mail: masato_hisakado@standardpoors.com, kj198276@sfc.keio.ac.jp and mori@sci.kitasato-u.ac.jp

Received 31 May 2006
Published 30 November 2006
Online at stacks.iop.org/JPhysA/39/15365

Abstract

We discuss a general method to construct correlated binomial distributions by imposing several consistent relations on the joint probability function. We obtain self-consistency relations for the conditional correlations and conditional probabilities. The beta-binomial distribution is derived by a strong symmetric assumption on the conditional correlations. Our derivation clarifies the 'correlation' structure of the beta-binomial distribution. It is also possible to study the correlation structures of other probability distributions of exchangeable (homogeneous) correlated Bernoulli random variables. We study some distribution functions and discuss their behaviours in terms of their correlation structures.

PACS number: 02.50.Cw

1. Introduction

Incorporation of correlation ρ into Bernoulli random variables X_i ($i = 1, 2, \ldots, N$) taking the value 1 with probability p and taking the value 0 with probability $1 - p$ has long history and have been widely discussed in a variety of areas of science, mathematics and engineering. Writing the expectation value of a random variable A as $\langle A\rangle$, the correlation ρ between X_i and X_j is defined as

$$\rho = \mathrm{Corr}(X_i, X_j) = \frac{\langle X_i X_j\rangle - \langle X_i\rangle\langle X_j\rangle}{\sqrt{\langle X_i\rangle(1 - \langle X_i\rangle)\langle X_j\rangle(1 - \langle X_j\rangle)}}. \qquad (1)$$

If there are no correlations between the random variables, the number n of the variables taking the value 1 obeys the binomial probability distribution $\mathrm{b}(N, p)$. The necessity of the correlation

ρ comes from the facts that there are many phenomena where dependency structures in the random events are crucial or are necessary for the explanation of experimental data.

For example, in biometrics, the teratogenic or toxicological effect of certain compounds was studied [1–3]. The interest resides in the number of affected fetuses or implantation in a litter. One-parameter models, such as the Poisson distribution and binomial distributions, provided poor fits to the experimental data. A two-parameter alternative to the above distributions, beta-binomial distribution (BBD), has been proposed [1, 2]. In the model, the probability p' of the binomial distribution $\mathrm{b}(N, p')$ is also a random variable and obeys the beta distribution $\mathrm{Be}(\alpha, \beta)$.

$$P(p') = \frac{p'^{\alpha-1}(1-p')^{\beta-1}}{\mathrm{B}(\alpha,\beta)}. \tag{2}$$

The resulting distribution has the probability function

$$P(n) = {}_N C_n \cdot \frac{\mathrm{B}(\alpha+n, N+\beta-n)}{\mathrm{B}(\alpha,\beta)}. \tag{3}$$

The mean μ and variance σ^2 of the BBD are

$$\mu = Np \qquad \text{and} \qquad \sigma^2 = Npq(1+N\theta)/(1+\theta), \tag{4}$$

where

$$p = \frac{\alpha}{\alpha+\beta}, \qquad q = 1-p = \frac{\beta}{\alpha+\beta} \qquad \text{and} \qquad \theta = \frac{1}{\alpha+\beta}. \tag{5}$$

θ is a measure of the variation in p' and is called as 'correlation level' [4]. The case of pure binomial distribution corresponds to $\theta = 0$. However, true 'correlation' of the BBD is given as

$$\rho = \frac{1}{\alpha+\beta+1}. \tag{6}$$

The derivation of the relation is straightforward. If we denote the sum of X_i as $S = \sum_{i=1}^{N} X_i$, we can write as $\langle X_i X_j\rangle = \langle S^2 - S\rangle/N(N-1)$ and $\langle X_i\rangle = \langle X_j\rangle = \langle S\rangle/N$. From equation (1) and the results for BBD, we obtain equation (6). We rewrite the variance σ^2 as

$$\sigma^2 = Npq + N(N-1)pq \cdot \rho. \tag{7}$$

In the area of computer engineering, in the context of the design of survivable storage system, the modelling of the correlated failures among storage nodes is a hot topic [4]. In addition to BBD, a correlated binomial model based on conditional failure probabilities has been proposed. The same kind of correlated binomial distribution based on conditional probabilities has also been introduced in financial engineering. There, credit portfolio modelling has been extensively studied [5, 6]. In particular, the modelling default correlation plays a central role in the pricing of portfolio credit derivatives, which are developed in order to manage the risk of joint default or the clustering of default. As a default distribution model for homogeneous (exchangeable) credit portfolio where the assets' default probabilities and default correlations are uniform and denoted as p and ρ, Witt has introduced a correlated binomial model based on the conditional default probabilities p_n [7]. Describing the defaulted (non-defaulted) state of ith asset by $X_i = 1$ $(X = 0)$ and the joint default probability function by $P(x_1, x_2, \ldots, x_N)$, p_n are defined as

$$p_n = \left\langle X_{n+1} \middle| \prod_{n'=1}^{n} X_{n'} = 1 \right\rangle. \tag{8}$$

Here $\langle A|B\rangle$ means the expectation value of a random variable A under the condition that B is satisfied. The expectation value of X_i signifies the default probability and the condition $\prod_{n'=1}^{n} X_{n'} = 1$ corresponds to the situation where the first n assets among N are defaulted. $p_0 = p$ and from the homogeneity (exchangeability) assumption, any n assets among N can be chosen in the n default condition $\prod_{n'=1}^{n} X_{n'} = 1$. X_{n+1} in equation (8) is also substituted by anyone which is not used in the n default condition.

In order to fix the joint default probability function completely, it is necessary to impose N conditions on them from the homogeneity assumption. Witt and the authors have imposed the following condition on the conditional correlations [7, 8]:

$$\mathrm{Corr}\left(X_{n+1}, X_{n+2}\,\middle|\,\prod_{n'=1}^{n} X_{n'} = 1\right) = \rho\exp(-\lambda n) \equiv \rho_n.$$

Here $\mathrm{Corr}(A, B|C)$ means that the correlation between the random variables A and B under the condition C is satisfied. From them, recursive relations for p_n are obtained and p_n are calculated as

$$p_n = 1 - (1-p)\prod_{n'=0}^{n-1}(1-\rho_{n'}).$$

The joint default probability function and the default distribution function $P_N(n)$ have been expressed with these p_n explicitly. However, the expression has many ± contributions and it is not an easy task to evaluate them for $N \geqslant 100$. In addition, the range of parameters p and ρ are also restricted and one cannot study the large correlation regime. Furthermore, for $p = 0.5$ case, the distribution does not have the Z_2 symmetry as $P_N(n) = P_N(N-n)$. The distribution has irregular shape and for some choice of parameters, it shows singular rippling.

In this paper, we propose a general method to construct correlated binomial models (CBM) based on the consistent conditions on the conditional probabilities and the conditional correlations. With the method, it is possible to study the correlation structure for any probability distribution function for exchangeable correlated Bernoulli random variables. The organization of the paper is as follows. In section 2, we introduce conditional probabilities p_{ij} and conditional correlations ρ_{ij} and show how to construct CBMs. We prove that the construction is self-consistent. In addition, in order to assure the probability conservation or the normalization, the conditional correlations and the probabilities should satisfy self-consistent relations. We also calculate the moments $\langle n^k\rangle$ of the model. In the course, we introduce a linear operator H which gives the joint probabilities in the 'binomial' expansion of $(p+q)^N$. Section 3 is devoted to some solutions of the self-consistent relations. We obtain the beta-binomial distribution (BBD) with strong symmetric assumptions on the conditional correlations. For other probability distribution functions which include the Witt's model and the distributions constructed by the superposition of the binomial distributions (Bernoulli mixture model), we calculate p_{ij} and ρ_{ij}. We study the probability distribution functions for these solutions from the viewpoint of their correlation structures ρ_{ij}. We conclude with some remarks and future problems in section 4.

2. Correlated binomial models and their constructions

In this section, we construct the joint probabilities and the distribution functions of CBMs. We introduce the following definitions. The first one is the products of X_i and $1 - X_j$ and

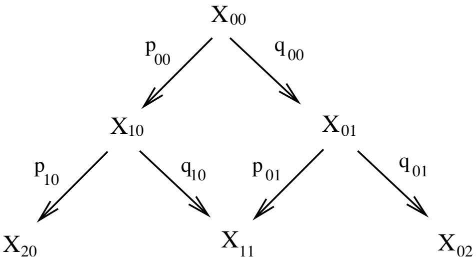

Figure 1. Pascal's triangle like representation of X_{ij} and p_{ij}, q_{ij} up to $i+j \leqslant 2$. $X_{00} = \langle 1 \rangle$, $X_{10} = \langle X_1 \rangle = p$, $X_{01} = \langle 1 - X_1 \rangle = 1 - p = q$ etc.

they include all observables of the model:

$$\Pi_{ij} = \prod_{i'=1}^{i} X_{i'} \prod_{j'=i+1}^{i+j} (1 - X_{j'}). \tag{9}$$

The following definitions are their unconditional and conditional expectation values (see figure 1):

$$X_{ij} = \langle \Pi_{ij} \rangle \tag{10}$$

$$p_{ij} = \langle X_{i+j+1} | \Pi_{ij} = 1 \rangle = \frac{X_{i+1j}}{X_{ij}} \tag{11}$$

$$q_{ij} = \langle 1 - X_{i+j+1} | \Pi_{ij} = 1 \rangle = \frac{X_{ij+1}}{X_{ij}}. \tag{12}$$

$X_{00} = 1$, $X_{10} = p$ and $X_{01} = 1 - p = q$. Furthermore, the relation $p_{ij} + q_{ij} = 1$ should hold for any i, j, because of the identity $\langle 1 | \Pi_{ij} = 1 \rangle = \langle X_{i+j+1} + (1 - X_{i+j+1}) | \Pi_{ij} = 1 \rangle = 1$. All informations are contained in X_{ij}. The joint probability $P(x_1, x_2, \ldots, x_N)$ with $\sum_{i'=1}^{N} x_{i'} = n$ is given by X_{nN-n} and the distribution function $P_N(n)$ is also calculated as

$$P_N(n) = {}_N C_n \cdot X_{nN-n}. \tag{13}$$

In order to estimate X_{ij}, we need to calculate the products of p_{kl} and q_{kl} from $(0, 0)$ to (i, j). As the path, we can choose anyone and the product must not depend on the choice. This property is guaranteed by the next condition on p_{ij} and q_{ij} as (see figure 2)

$$q_{i+1j} \cdot p_{ij} = p_{ij+1} \cdot q_{ij} = \frac{X_{i+1j+1}}{X_{ij}}. \tag{14}$$

In order for p_{ij} and q_{ij} to satisfy these conditions, we introduce the following conditional correlations:

$$\mathrm{Corr}(X_{i+j+1}, X_{i+j+2} | \Pi_{ij} = 1) = \rho_{ij}. \tag{15}$$

We set $\rho_{00} = \rho$. $(1 - X_{i+j+1})$ and $(1 - X_{i+j+2})$ are also correlated with the same strength and the following relations hold.

$$\mathrm{Corr}((1 - X_{i+j+1}), (1 - X_{i+j+2}) | \Pi_{ij} = 1) = \rho_{ij}. \tag{16}$$

From these relations, we obtain the recursive relations for p_{ij} and q_{ij} as

$$\begin{aligned} p_{i+1j} &= p_{ij} + (1 - p_{ij})\rho_{ij} \\ q_{ij+1} &= q_{ij} + (1 - q_{ij})\rho_{ij}. \end{aligned} \tag{17}$$

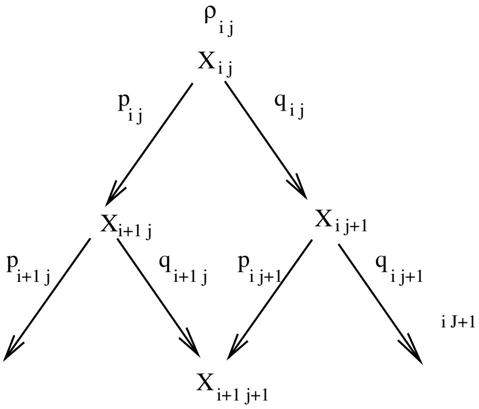

Figure 2. Proof of the commutation relation $q_{i+1j} \cdot p_{ij} = p_{ij+1} \cdot q_{ij}$.

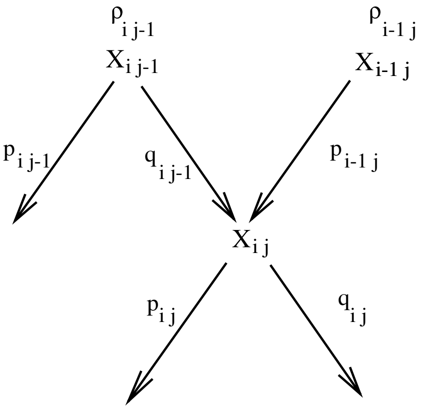

Figure 3. Picture for the $p_{ij} + q_{ij} = 1$ condition.

If we assume the identity $p_{ij} + q_{ij} = 1$, we obtain $q_{ij} = 1 - p_{ij}, q_{i+1j} = 1 - p_{i+1j} = (1 - p_{ij})(1 - \rho_{ij})$ and $p_{ij+1} = 1 - q_{ij+1} = p_{ij}(1 - \rho_{ij})$. Then $q_{i+1j} \cdot p_{ij} = p_{ij+1} \cdot q_{ij} = p_{ij}(1 - p_{ij})(1 - \rho_{ij})$ holds and we see that the above consistency relation (14) does hold.

The remaining consistency relations or the probability conservation identity is $p_{ij} + q_{ij} = 1$. We prove the identity by the inductive method (see figure 3). For $i = j = 0$, the identity holds trivially as $p_{00} + q_{00} = p + q = 1$. For $j = 0$ or $i = 0, q_{i0}$ and p_{0j} are calculated as $q_{i0} = 1 - p_{i0}$ and $p_{0j} = 1 - q_{0j}$ and the identity also holds trivially. Then we assume $p_{ij-1} + q_{ij-1} = 1$ and prove the identity $p_{ij} + q_{ij} = 1$. From the recursive equations (17) on p_{ij} and q_{ij}, we have the following relations.

$$1 = p_{ij} + q_{ij} = p_{i-1j} + (1 - p_{i-1j})\rho_{i-1j} + (1 - p_{ij-1}) + p_{ij-1}\rho_{ij-1}. \tag{18}$$

For the identity to be satisfied, the conditional correlation ρ_{ij-1} and ρ_{i-1j} must satisfy the following relations.

$$p_{i-1j} - p_{ij-1} = -(1 - p_{i-1j})\rho_{i-1j} - p_{ij-1}\rho_{ij-1}. \tag{19}$$

If the conditional correlations ρ_{ij} are fixed so as to satisfy the relations, the model becomes self-consistent. In other words, it guarantees the normalization of the resulting probability distribution.

We estimate the moments of CBM. For the purpose, we introduce following operators H and D_k. The former one is a linear operator H which maps polynomial in p, q to joint probabilities $\in R$. By its linearity, we only need to fix its action on monomial $p^i q^j$ as

$$H[p^i q^j] = p_{00} p_{10} \cdots p_{i-10} q_{i0} q_{i1} \cdots q_{ij-1}. \tag{20}$$

The joint probability X_{nN-n} is expressed as $X_{nN-n} = H[p^n q^{N-n}]$. Here we choose the far left path from (0, 0) to $(n, N-n)$ on the Pascal's triangle (see figure 1). The action of H on the binomial expansion $(p+q)^N = 1^N$ can be interpreted as the probability distribution and its normalization condition:

$$1 = H[1^N] = H[(p+q)^N] = \sum_{n=0}^{N} {}_N C_n \cdot H[p^n q^{N-n}] = \sum_{n=0}^{N} {}_N C_n \cdot X_{nN-n}. \tag{21}$$

In order to calculate the moments of CBM, it is necessary to put n^k in the above summation. Instead, we will put $n(n-1)(n-2)\cdots(n-k+1)$ and introduce the following differential operators D_k:

$$D_k = \sum_{0 \leqslant i_1, i_2, \ldots \leqslant N-1}^{i_1 \neq i_2, i_1 \neq i_3, \ldots, i_{k-1} \neq i_k} p_{i_1 0} p_{i_2 0} \cdots p_{i_k 0} \frac{\partial^k}{\partial p_{i_1 0} \partial p_{i_2 0} \cdots \partial p_{i_k 0}}. \tag{22}$$

The action of D_k on X_{nN-n} for $n \geqslant k$ is

$$D_k X_{nN-n} = n(n-1)(n-2)\cdots(n-k+1) X_{nN-n}. \tag{23}$$

On the other hand, the same expression can be obtained as

$$\begin{aligned} H\left[p^k \frac{\mathrm{d}^k}{\mathrm{d}p^k} p^n q^{N-n}\right] &= H[n(n-1)(n-2)\cdots(n-k+1) p^n q^{N-n}] \\ &= n(n-1)(n-2)\cdots(n-k+1) X_{nN-n}. \end{aligned} \tag{24}$$

This relation defines the action of D_k on the operator H with any polynomial $f(p, q)$ as

$$D_k H[f(p,q)] = H\left[p^k \frac{\mathrm{d}^k}{\mathrm{d}p^k} f(p,q)\right]. \tag{25}$$

The calculation of the expectation value of $n(n-1)\cdots(n-k+1)$ is performed by the action of operator D_k on the binomial expansion of $H[1^N] = H[(p+q)^N]$:

$$D_k H[(p+q)^N] = \sum_{n=0}^{N} {}_N C_n \cdot D_k X_{nN-n}. \tag{26}$$

The right-hand side is nothing but the expectation value $\langle n(n-1)(n-2)\cdots(n-k+1)\rangle$. The left-hand side is calculated by using equation (25) as

$$\begin{aligned} D_k H[(p+q)^N] &= H\left[p^k \frac{\mathrm{d}^k}{\mathrm{d}p^k} (p+q)^N\right] \\ &= N(N-1)(N-2)\cdots(N-k+1) H[p^k (p+q)^{N-k}] \\ &= N(N-1)(N-2)\cdots(N-k+1) H[p^k] \\ &= N(N-1)(N-2)\cdots(N-k+1) p_{00} p_{10} p_{20} \cdots p_{k-10}. \end{aligned} \tag{27}$$

We obtain the relation

$$\langle n(n-1)(n-2)\cdots(n-k+1)\rangle = N(N-1)(N-2)\cdots(N-k+1)p_{00}p_{10}p_{20}\cdots p_{k-10}. \tag{28}$$

From the relation, we can estimate the moments of CBM.

3. Beta-binomial distribution and other solutions

In the previous section, we have derived self-consistent equations for p_{ij} and ρ_{ij}. They are summarized as

$$p_{i+1j} = p_{ij} + (1-p_{ij})\rho_{ij} \tag{29}$$

$$p_{ij+1} = p_{ij} - p_{ij}\rho_{ij} \tag{30}$$

$$p_{i-1j} - p_{ij-1} = -(1-p_{i-1j})\rho_{i-1j} - p_{ij-1}\rho_{ij-1}. \tag{31}$$

In this section, we show several solutions to these equation. We note, if one knows joint probabilities X_{ij}, from the definitions for p_{ij} and q_{ij}, we can estimate p_{ij}. Then ρ_{ij} are estimated from the recursive equation (29). In addition, we interpret the behaviours of the solutions from the viewpoint of correlation structures.

3.1. Beta-binomial distribution

In order to solve the above relations on ρ_{ij} and p_{ij}, we use the symmetry viewpoint. For $p = \frac{1}{2}$ case, the model should have particle–hole duality between X and $1-X$ or Z_2 symmetry. Then $\rho_{ij} = \rho_{ji}$ should hold. We put stronger assumption that for any p, the system has the Z_2 symmetry and ρ_{ij} depends on i, j only through the combination $n = i + j$. With a suitable choice of indexes $i \to i+1$ and $j = n-i$, equation (31) reduces to

$$p_{in-i} - p_{i+1n-i-1} = \rho_n(-1 + p_{in-i} - p_{i+1n-i-1}). \tag{32}$$

From this relation, we see that p_{ij} with the same $n = i + j$ consist an arithmetic sequence with the common difference Δ_n.

$$p_{i+1n-i-1} - p_{in-i} = \Delta_n. \tag{33}$$

Δ_n satisfy the following equation:

$$\Delta_n = \rho_n(1+\Delta_n). \tag{34}$$

ρ_n can be solved with Δ_n as

$$\rho_n = \frac{\Delta_n}{1+\Delta_n}. \tag{35}$$

From relation (29) for p_{ij}, we obtain the following recursive relation for ρ_n as,

$$\rho_n = \frac{\Delta_n}{1+\Delta_n} = \frac{\Delta_{n-1}(1-\rho_{n-1})}{1+\Delta_{n-1}(1-\rho_{n-1})} = \frac{\rho_{n-1}}{1+\rho_{n-1}}. \tag{36}$$

The explicit form for ρ_n and Δ_n are

$$\rho_n = \frac{\rho}{1+n\rho} \qquad \text{and} \qquad \Delta_n = \rho_{n-1}. \tag{37}$$

Then p_{ij} and q_{ij} can be obtained explicitly and the results are

$$p_{ij} = p_{i+j0} - j\Delta_{i+j} = \frac{p(1-\rho)+i\rho}{1+(i+j-1)\rho} \tag{38}$$

$$q_{ij} = 1 - p_{ij} = \frac{q(1-\rho)+j\rho}{1+(i+j-1)\rho}. \tag{39}$$

X_{nN-n} are then obtained by taking the products of these conditional probabilities from $(0, 0)$ to $(n, N-n)$,

$$X_{nN-n} = \prod_{i=0}^{n-1} p_{i0} \prod_{j=0}^{N-n-1} q_{nj}. \tag{40}$$

Putting the above results for p_{ij} and q_{ij} into them, we obtain

$$X_{nN-n} = \frac{\prod_{i=0}^{n-1}(p(1-\rho)+i\rho)\prod_{j=0}^{N-n-1}(q(1-\rho)+j\rho)}{\prod_{k=0}^{N-1}(1+(k-1)\rho)}. \tag{41}$$

Here $q = 1-p$. By multiplying the binomial coefficients ${}_NC_n$, we obtain the distribution function $P_N(n)$ as

$$P_N(n) = {}_NC_n \cdot X_{nN-n}. \tag{42}$$

This distribution is nothing but the beta-binomial distribution function (see equation (3)) with suitable replacements $(p, \rho) \leftrightarrow (\alpha, \beta)$.

3.2. Moody's correlated binomial model

In the original work by Witt, he assumed $\rho_{i,0} = \rho$ for all i [7]. We call this model as Moody's correlated binomial (MCB) model. The above consistent equations are difficult to solve and the available analytic expressions are those for p_{i0} as $p_{i0} = 1-(1-p)(1-\rho)^i$. With the result, we only have a formal expression for X_{ij} as

$$\begin{aligned} X_{ij} = \langle \Pi_{ij} \rangle &= \left\langle \prod_{i'=1}^{i} X_{i'} \prod_{j'=i+1}^{i+j} (1-X_{j'}) \right\rangle \\ &= \sum_{k=0}^{j} (-1)^k {}_jC_k \left\langle \prod_{i'=1}^{i+k} X_{i'} \right\rangle = \sum_{k=0}^{j} (-1)^k {}_jC_k \cdot p_{i+k0}. \end{aligned} \tag{43}$$

With this expression, it is possible to estimate p_{ij}, q_{ij} and ρ_{ij} from their definitions. However, equation (43) contains ${}_jC_k(-1)^k$ and as N becomes large, it becomes difficult to estimate them. With the above choice for $\rho_{i0} = \rho$, it is possible to set $N = 30$. If ρ_{i0} damps as $\exp(-\lambda i)$ with some positive λ, we can set at most $N = 100$ for small values of ρ and p.

3.3. Mixed binomial models: Bernoulli mixture models

Bernoulli mixture model with some mixing probability distribution function $f(p)$, the expression for the joint probability function X_{ij}, is calculated with

$$X_{ij} = \langle \Pi_{ij} \rangle = \int_0^1 \mathrm{d}p\, f(p) p^i (1-p)^j. \tag{44}$$

If we use the beta distribution for $f(p)$, we obtain equation (41). However, this does not mean that it is trivial to solve the consistent equations with the assumption $\rho_{ij} = \rho_{i+j}$ and obtain the BBD. The consistent equations completely determine any correlated binomial distribution for exchangeable Bernoulli random variables. Every correlated binomial distribution obeys the relations. With the assumption $\rho_{ij} = \rho_{i+j}$, we are automatically led to the BBD. That is, the probability distribution with the symmetry $\rho_{ij} = \rho_{i+j}$, we prove that it is the BBD. No other probability distribution has the symmetry.

Here we consider the relation between CBM and Bernoulli mixture model. According to De Finetti's theorem, the probability distribution of any infinite exchangeable Bernoulli random variables can be expressed by a mixture of the binomial distribution [10]. CBM in the $N \to \infty$ limit should be expressed by such a mixture. From equation (44), we have the relation $P(x_1 = 1, x_2 = 1, \ldots, x_k = 1) = X_{k0} = \int f(p)p^k \,\mathrm{d}p$. X_{k0} is expressed as $X_{k0} = p_{00}p_{10}\cdots p_{k-10}$, we have a correspondence between the moments of $f(p)$ and a CBM. That is, if one knows p_{i0} for any i, we know the mixing function $f(p)$ and vice versa. This correspondence shows the equivalence of CBM and the Bernoulli mixture model in the large N limit. But CBMs with finite N can describe probability distribution more widely. In the Bernoulli mixture model, the variance of p is positive and the correlation ρ cannot be taken negative. In CBM, we can set ρ negative for small system size N. In addition, CBM is useful to construct the probability distribution and discuss about the correlation structure. Particularly we can understand the symmetry of the solution. For example, we want to have Z_2 symmetry distributions. In the Bernoulli mixture model, we need to impose on $f(p)$ as

$$\int_0^1 f(p)(p-0.5)^{2k+1}\,\mathrm{d}p = 0, \tag{45}$$

where $k = 1, 2, \ldots$. On the other hand, in CBM, we only need to seek a solution with $p_{ii} = q_{ii} = \frac{1}{2}$. This simple constraint is useful in the construction and in the parameter calibration of CBMs.

As other mixing functions $f(p)$, we consider the cases which correspond to the long-range Ising model with some strength of magnitude of correlation $\rho > 0$. It has some correlation only in the regime where the probability distribution for the magnetization $p(m)$ has two peaks at m_1, m_2 for $T < T_c$ [9]. If the system size N is large enough, the distribution can be approximated with the superposition of two binomial distributions. If we take $N \to \infty$ for $T < T_c$, the system loses its ergodicity and the phase space breaks up into two space with $m > 0$ and $m < 0$ [11] and the correlation disappears. Even if there appears two peaks in $p(m)$, only one of them represents the real equilibrium state.

The precise values of m_1 and m_2 depend on the model parameters, we consider the cases which correspond to $p = 0.5$ (Z_2 symmetric case) and $p \simeq 0$. For the Z_2 symmetric case, there is no external field and $m_1 = -m_2$ holds. Between the Bernoulli random variable X and the Ising spin variable S, there exists a mapping $X = \frac{1}{2}(1-S)$. $f(p)$ has two peaks at p and $q = 1-p$ with the same height. On the other hand, for $T \simeq 0$ and infinitely weak positive external field case $\sim O\left(\frac{1}{N}\right)$, $p(m)$ has one tall peak at $m_1 \simeq 1$ and another short peak at $m_2 \simeq -1$. In the language of the Bernoulli random variable case, $f(p)$ has a tall peak at $p' = p'' \simeq 0$ and a short peak at $p' \simeq 1$. We consider the following mixing functions and call them two-binomial models.

- $f(p') = \frac{1}{2}\delta(p'-p) + \frac{1}{2}\delta(p'-q)$ with $q = 1-p$.
  This mixing function corresponds to the long-range Ising model with Z_2 symmetry and $\rho > 0$. X_{ij} are given as

$$X_{ij} = \tfrac{1}{2}(p^i q^j + p^j q^i). \tag{46}$$

  p_{ij} and ρ_{ij} are calculated easily as

$$p_{ij} = \frac{p^{i+1}q^j + p^j q^{i+1}}{p^i q^j + p^j q^i} \tag{47}$$

$$\rho_{ij} = \frac{p^{i+j}q^{i+j}(p-q)^2}{p^{i+j}q^{i+j}(p^2+q^2)+qp(p^{2i}q^{2j}+q^{2i}p^{2j})}. \tag{48}$$

This solution has the Z_2 symmetry $\rho_{ij} = \rho_{ji}$.

- $f(p') = \frac{p^k}{p^k+q^k}\delta(p'-p) + \frac{q^k}{p^k+q^k}\delta(p'-q)$ with $q = 1-p$.
  This is the modified version of the above solution with a parameter $k = 0, 1, \ldots$. If we set $k = 0$, it is nothing but the above solution. X_{ij} are given as

$$X_{ij} = \frac{1}{p^k+q^k}(p^i q^j p^k + p^j q^i q^k). \tag{49}$$

p_{ij} and ρ_{ij} are

$$p_{ij} = \frac{p^{i+k+1}q^j + p^j q^{i+k+1}}{p^{i+k}q^j + p^j q^{i+k}} \tag{50}$$

$$\rho_{ij} = \frac{p^{i+j+k}q^{i+j+k}(p-q)^2}{p^{i+j+k}q^{i+j+k}(p^2+q^2)+qp(p^{2i+2k}q^{2j}+q^{2i+2k}p^{2j})}. \tag{51}$$

If we denote $C_1 = \frac{p^k}{p^k+q^k}, C_2 = \frac{q^k}{p^k+q^k}$, then the mixing function becomes $f(p') = C_1\delta(p'-p)+C_2\delta(p'-q)$. This solution may look trivial. One obtain this solution using the parallel shift of the above solution (46). We replace X_{ij} with X_{i+kj} in equation (46) and obtain the solution. Such a parallel shift may give birth to another solution, we would like to note it here.

- $f(p') = (1-\alpha)\delta(p'-p'') + \alpha\delta(p'-1)$.
  This mixing function corresponds to the long-range Ising model without Z_2 symmetry, $\langle S_i\rangle \simeq 1$ and $\rho > 0$. We call the model as Binomial plus (B+) model, because it is a binomial distribution plus one small peak at $n = N$. Between p, ρ and p'', α, we have the relations

$$p = \alpha + (1-\alpha)p'' \qquad \text{and} \qquad \rho = \frac{\alpha(1-p'')}{\alpha+(1-\alpha)p''} \tag{52}$$

and

$$\alpha = \frac{\rho p}{1-p+\rho p}. \tag{53}$$

X_{ij} are given as

$$X_{ij} = (1-\alpha)p''^i(1-p'')^j + \alpha\delta_{j,0}. \tag{54}$$

p_{ij} and ρ_{ij} are calculated easily as

$$p_{i0} = \frac{\alpha+(1-\alpha)p''^{i+1}}{\alpha+(1-\alpha)p''^i} \qquad \text{and} \qquad p_{ij} = p'' \quad \text{for} \quad j \neq 0 \tag{55}$$

and

$$\rho_{i0} = \frac{\alpha(1-p'')}{\alpha+(1-\alpha)p''^{i+1}} \qquad \text{and} \qquad \rho_{ij} = 0 \quad \text{for} \quad j \neq 0. \tag{56}$$

3.4. Correlation structures of the solutions

In this subsection, we study the relations between probability distributions and correlation structure. Figure 4 shows the probability distribution profiles for three correlated models,

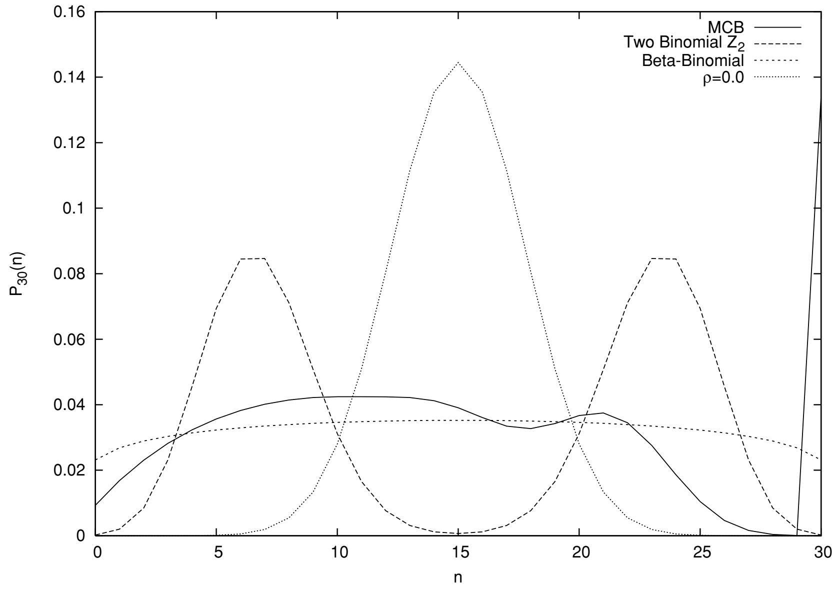

Figure 4. Probability distribution $P_{30}(n)$ for $p = 0.5, \rho = 0.3$ and $N = 30$. We show 3 distributions, MCB (solid line), beta-binomial (dotted line) and two-binomial (thin dotted line). We also show a binomial distribution ($\rho = 0.0$) for comparison.

MCB, BBD and two-binomial models. We set $p = 0.5$, $\rho = 0.3$ and $N = 30$. We also shows the pure binomial distribution for comparison. The former three curves have the same p and ρ, however their profiles are drastically different. Two-binomial model with Z_2 symmetry has two peaks and their overlap decreases as N increases. At the thermodynamic limit $N \to \infty$, the overlap disappears and the system loses its ergodicity. The long-range Ising models shows spontaneous symmetry (SSB) breaking of the Z_2 symmetry. On the other hand, the BBD's profile is broad and even if we set $N \to \infty$, we obtain the beta distribution and the shape is almost unchanged. That is, the BBD system does not show SSB and it maintains its Z_2 (particle–hole) symmetry at $p = 0.5$.

The profile of MCB model is peculiar. It is not symmetric and shows singular rippling. The origin for the rippling can be understood from the inspection of its correlation structure. Figure 5 shows the correlation structures for the above three models. The parameters are equal and we show ρ_{i30-i}. In contrast to the BBD's correlation, which is constant with $i + j$ fixed, the correlations for MCB have a sharp peak at $i = 30$ and show strong rippling structure. The curve is not symmetric and the distortion is reflected in the shape of its probability distribution. On the other hand, the correlation curve for two-binomial distribution has a strong peak at $i = \frac{N}{2}$ and it is much different from the BBD's correlation curve. This strong peak and rapid decay may be reflected in the decomposition of the probability distribution. However, we have not yet understood the relation well.

Figure 6 shows the probability distribution for MCB, BBD and B+ models. We set $p = 0.1$, $\rho = 0.3$ and $N = 30$. We also show the pure binomial distribution for comparison. MCB and BBD have almost the same bulk shape; however, MCB has a small peak at $n = 30$. B+ has more strong peak at $n = 30$ and its bulk shape can be obtained by a small left shift of the pure binomial distribution $p = 0.1$. These profile differences are reflected in their correlation structures (see figure 7). It shows the correlation structures for the above three

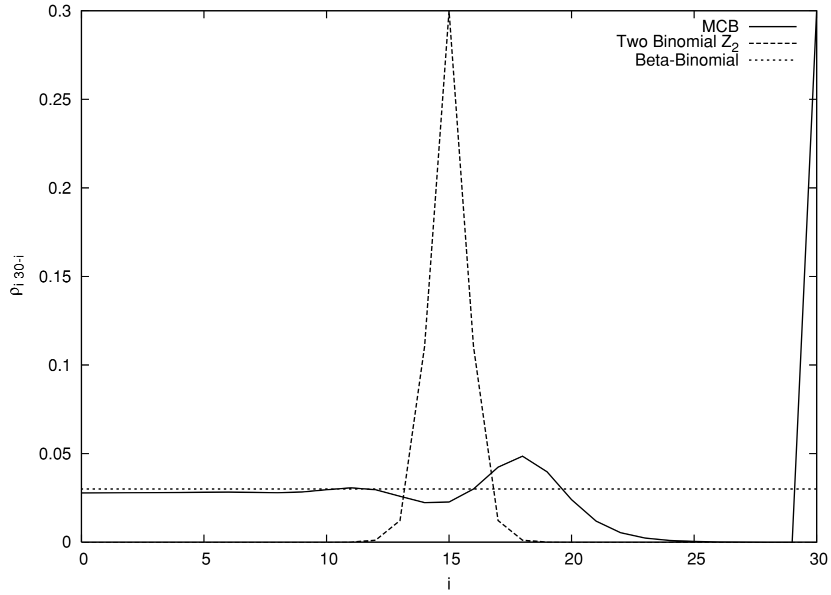

Figure 5. Correlation ρ_{i30-i} for MCB (solid line), BBD (thin dotted line) and two-binomial (dotted line) models. We set $\rho = 0.3$ and $p = 0.5$ as in the previous figure.

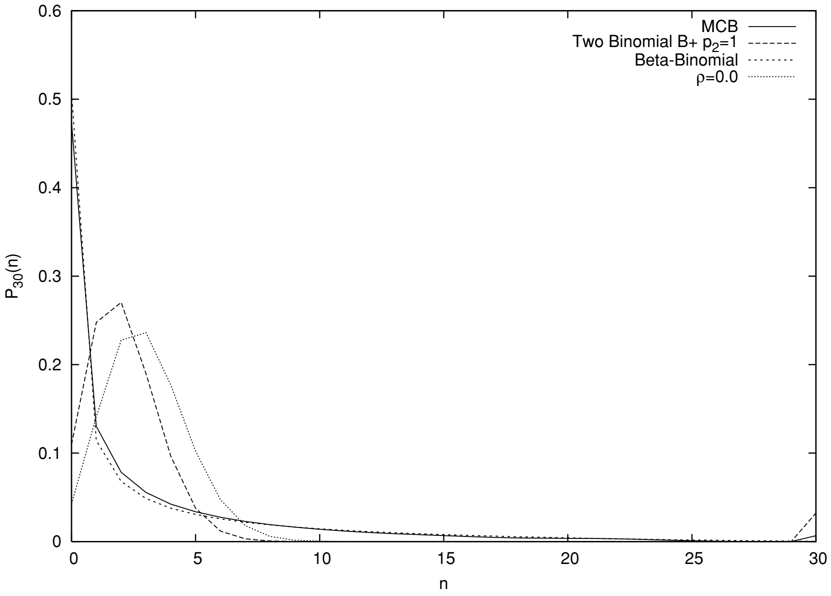

Figure 6. Probability distribution $P_{30}(n)$ for $p = 0.1, \rho = 0.3$ and $N = 30$. We show three distributions, MCB (solid line), beta-binomial (dotted line) and B+ (thin dotted line). We also show a binomial distribution ($\rho = 0.0$) for comparison.

models. The parameters are equal as in the previous figure. Contrary to the constant BBD structure, MCB and B+ models have a peak at $i = 30$. MCB has a small and B+ has a tall peak and the difference is reflected in the size of their tall peak of the probability distributions.

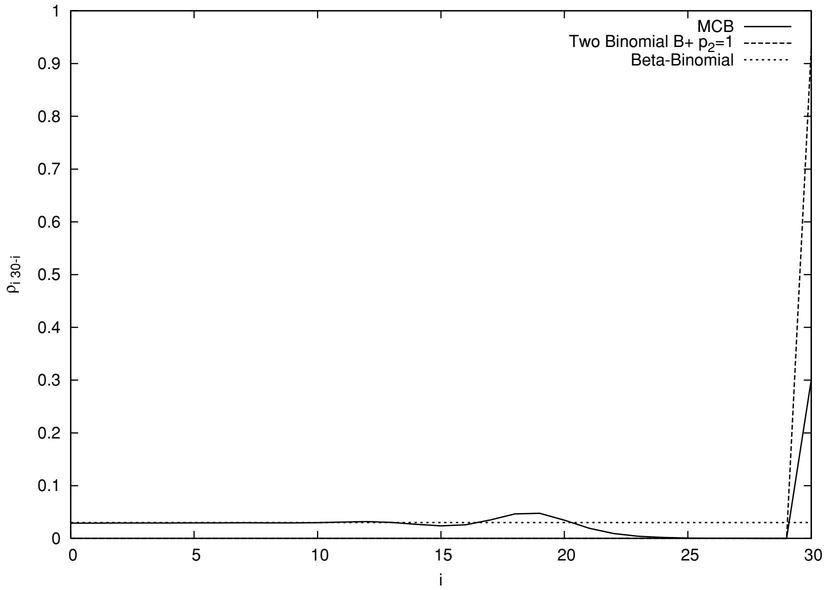

Figure 7. Correlation ρ_{i30-i} for MCB (solid line), BBD (thin dotted line) and B+ (dotted line) models. We set $\rho = 0.3$ and $p = 0.1$.

4. Concluding remarks and future problems

In this paper, we show a general method to construct correlated binomial models. We also estimate their moments. Our method includes Witt's model and the BBD. In addition, with the consistent equations on p_{ij} and ρ_{ij}, it is possible to prepare correlated binomial distributions with any choice for ρ_{i0} or p_{i0}. Of course, the resulting distribution function should be non-negative, 'any' should be taken with some care. In addition, from the joint probabilities X_{ij}, it is possible to estimate p_{ij} and ρ_{ij}. We can see the detailed structure of the system with any distribution function. In the work [4], the conditional strange failure probabilities p_{i0} were studied. Some recursive relations on p_{i0} were proposed and the resulting conditional probabilities p_{i0} were compared with real data on server networks. We note that p_{i0} can be freely changed and it may be possible to make a good fitting with data. However, if the correlation structure ρ_{ij} becomes too complex and it shows oscillation, such a modelling may be over-fitting.

At last, we make comments about future problems. The first one is to seek another interesting solution to equations (29)–(31) about ρ_{ij} and p_{ij}. In this paper, we have assumed strong symmetry in ρ_{ij} in the derivation of the BBD. For any value of p, we have assumed Z_2 symmetry $\rho_{ij} = \rho_{ji}$. Furthermore, we have assumed stronger constraint that ρ_{ij} depends on i, j only through the combination $i + j$. The consistent relation is then solved easily and we get the BBD. However, we think that the correlated binomial distribution space is rich and there may exist other interesting solutions. We discuss some simple solutions which are superpositions of two binomial distribution. They try to mimic the long-range Ising model in the large N limit and $\rho > 0$ [9]. A simple seamless solution for the consistent relations which corresponds to the long-range Ising model may exist. Taking the continuous limit of the consistent relations and studying their solution is also an interesting problem. The solution space may become narrow, however differential equations are more tractable than the recursion relations. There should exist the beta distribution and the superposition of delta-functions, which are the continuous limits of the simple solutions presented here.

The second problem is the generalization of the present method. In this paper, we have assumed that the Bernoulli random variables are all exchangeable. If one considers to apply the correlated binomial model to the real world, such an idealization should be relaxed. One possibility is the inhomogeneity in p and the other is the inhomogeneity in ρ. The first step is to add one other Bernoulli random variable Y to N exchangeable variable system. This $N+1$ system case has been treated in [8], it seems much difficult to introduce the self-consistent equations in the present context. However, such a generalization may lead us to find new probability distribution functions, we believe that it deserves for extensive studies.

References

[1] Griffiths D A 1973 *Biometrics* **29** 637
[2] Williams D A 1975 *Biometrics* **31** 949
[3] Kupper L L and Haseman J K 1978 *Biometrics* **34** 69
[4] Bakkaloglu M *et al* 2002 *Technical Report* CMU-CS-02-129 Carnegie Mellon University
[5] Schönbucher P J 2003 *Credit Derivatives Pricing Models: Model, Pricing and Implementation* (New York: Wiley)
[6] Frey R and McNeil A J 2003 *J. Risk* **6** 59
[7] Witt G 2004 *Moody's Correlated Binomial Default Distribution* (Moody's Investors Service)
[8] Mori S, Kitsukawa K and Hisakado M 2006 Moody's correlated binomial default distributions for inhomogeneous portfolios (*Preprint* physics/0603036)
[9] Kitsukawa K, Mori S and Hisakado M 2006 Evaluation of Tranche in Securitization and long-range Ising model *Physica* A 368(2006)191
[10] Finetti De 1974-5 *Theory of Probability* (New York: Wiley)
[11] Goldenfeld N 1992 *Lectures on Phase Transitions and the Renormalization Group* (Reading, MA: Addison-Wesley)