Journal of Statistical Research *ISSN 0256 - 422 X*
2009, Vol. 43, No. 2, pp. 3-15
Bangladesh

A SIMPLE GRAPHICAL METHOD TO CHECK DEPENDENCE STRUCTURE USING COPULA

Eun-Joo Lee and Sarah Dial
Department of Mathematics, Millikin University, Decatur, IL 62522, USA
Email: elee@millikin.edu, sdial@millikin.edu

Leann Stuber and Seung-Hwan Lee
Department of Mathematics, Illinois Wesleyan University, Bloomington, IL 61702, USA
Email: lstuber@iwu.edu, slee2@iwu.edu

SUMMARY

Checking dependence structure between variables is important when modeling multivariate data. In this paper, we investigate copulas, functions that provide a flexible methodology for modeling multivariate dependence. In particular, copulas are useful in describing the dependence structure of extreme values in the tails. Gauss, Student t- and Cauchy copulas are considered, among others. We also investigate a method for the choice of copulas. Based on the chosen copula, we demonstrate a simple graphical method with data from a study on multiple myeloma, particularly for analysis of the tail dependence.

Keywords and phrases: Copulas; Dependence measures; Simulation

AMS Classification: 62E10,62-09,62P10

1 Introduction

An important problem in statistical modeling is to check the dependence between variables (Lehmann, 1966). The dependence measure most frequently used is Pearson's correlation coefficient, due to its easy computation and simple interpretation. However, Pearson's correlation coefficient has several drawbacks as a measure of dependence (Embrechts, McNeil and Straumann, 1999; Frees and Valdez, 1998; Schweizer and Wolff, 1981, Schweizer, 1991). Pearson's coefficient only measures linear dependence in terms of a single number. In practice, there are many occasions that have a non-linear dependence between variables. In addition, Pearson's correlation coefficient is strongly affected by extreme values and not invariant under non-linear increasing transformations of random variables. For these reasons, the linear correlation coefficient may not be an ideal measure of dependence. An alternative to the linear correlation coefficient is a copula function which overcomes the above problems as a measure of the dependence structure. The concept of copulas was introduced by Sklar

(1959). Copulas have been frequently used in the area of survival analysis (Klein and Zheng, 1995), risk management and financial applications (Breymann et al., 2003; Embrechts et al., 1999, 2003), all of which heavily focus on the extreme value analysis. Copulas are functions linking multivariate distributions to their marginal distributions, where the marginal distributions are uniform over the interval [0,1]. There are several advantages to using copulas, including their ability to capture both linear and non-linear dependence between variables. Also, the dependence captured by a copula is invariant under monotone transformations. More importantly, copulas measure the amount of dependence between random variables in the tails, so they are useful in analyzing the dependence structure of extreme values in these regions. For example, in survival analysis, one may be interested in the event that treatment on some disease either exceeds or falls below given levels at the early or late stage. Excellent reviews on copulas can be found in Nelson (1999).

In this paper, we study elliptical copulas, like the Gaussian copula, t-copula, and Cauchy copula. It is important to select appropriate copulas that may fit best for data. This is because poorly chosen copulas may lead to some undesirable consequences. For example, Gaussian copula may understate the magnitude of tail dependence of variables when the extreme events are actually highly dependent. Based on the best copula that is chosen from the L^2 norm criterion, similar to Shim et al. (2009), we demonstrate a simple graphical method with data from a study on multiple myeloma, particularly for analysis of the tail dependence. This work is organized as follows: dependence measures are discussed in section 2, and copulas are given in section 3. Section 4 presents simulations using copulas and real data analysis, then section 5 concludes this paper.

2 Dependence Measures

Let $X_i, i = 1, \ldots, n$, be random variables that have a marginal distribution, $F_i(x)$, defined as $P(X_i \leq x)$. The dependence between the real-valued random variables $X_1, \ldots, X_n$ is described by their joint distribution function

$$F(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

The conventional method to measure the dependence between the random variables is to use Pearson's linear correlation coefficient defined as, for an arbitrary pair (X_l, X_m),

$$\rho_{X_l, X_m} = \frac{\mathrm{Cov}(X_l, X_m)}{\sigma_{X_l} \sigma_{X_m}},$$

where $\mathrm{Cov}(X_l, X_m)$ represents the covariance of (X_l, X_m), and both σ_{X_l} and σ_{X_m} are the standard deviations of X_l and X_m, respectively. Pearson's correlation coefficient ranges from -1 to 1 respectively representing a perfect negative linear correlation or a perfect positive linear correlation between the variables.

However, Pearson's correlation coefficient is not a complete description of the dependence structure even when there is a straight-line relationship between two variables (Kumar and

Shoukri, 2007). It can be influenced by extreme values and is not invariant under non-linear monotonic increasing transformations of random variables. Other measures of dependence are Spearman's rank correlation and Kendall's rank correlation. They measure the association only in terms of ranks, while Pearson's correlation is data-dependent. Therefore both Spearman's and Kendall's rank correlations may be used as measures of the degree of monotonic dependence between random variables, while Pearson's correlation coefficient measures the degree of linear dependence only. Moreover, those rank based measures are invariant under any monotonic transformations, which is an advantage over Pearson's correlation. Using the rank-based dependence measures computed from the data, the copula parameter, specified later, is estimated.

Nevertheless, in practice, measuring dependence between variables with a single number like linear or rank correlation is somewhat limited since such measures are unlikely to completely capture extreme tail dependence. This may happen in areas dealing with insurance risks and life times, for instance. In addition, fitting the aforementioned multivariate joint probability distribution to data is complicated, especially when a large number of variables are involved. It is very unlikely to accurately estimate such a joint distribution function.

One way to overcome such problems is to use what is called a copula. Both linear and non-linear dependence between variables can be detected through a copula. More importantly, it establishes tail dependence among variables. Specifically, a copula describes the amount of dependence in the upper or lower tails, so it can be used in analyzing the dependence structure among extreme values in the tails. The coefficients of tail dependence describe these dependence measures in the tails. For the bivariate distribution of X_1 and X_2, the the coefficients of lower and upper tail dependences of X_1 and X_2 are defined as

$$\lim_{u\to 0+} P(X_2 > F_2^{-1}(u)|X_1 > F_1^{-1}(u))$$
$$\lim_{u\to 1-} P(X_2 > F_2^{-1}(u)|X_1 > F_1^{-1}(u))$$

respectively, provided that each limit exists in the interval . The coefficient of tail dependence given above will be discussed later in association with copula functions.

3 Copulas

Definition 3.1. A copula is a function $C : [0,1]^n \to [0,1]$ which has the following conditions:

C1. $C(u_1, \ldots, u_{i-1}, 0, u_{i+1}, \ldots, u_n) = 0$;
$C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i$ for all $i \in \{1, \ldots, n\}$, $u_i \in [0,1]$;

C2. $C(u_1, \ldots, u_k)$ is nondecreasing in each component u_i.

C3. For all $(u_{11}, \ldots, u_{n1}), (u_{12}, \ldots, u_{n2}) \in [0,1]^n$ with $u_{i1} \leq u_{i2}$, we have
$\sum_{i_1=1}^{2} \cdots \sum_{i_n=1}^{2} (-1)^{i_1+\ldots+i_n} C(u_{1i_1}, \ldots, u_{ni_n}) \geq 0$.

Note that C1 indicates the existence of the uniform distributions as the marginal distributions. C2 and C3 represent the fact that a copula is the distribution function of a random

vector. To summarize, a copula is the distribution function of a random vector with uniform $(0, 1)$ marginal distribution functions.

The idea of extracting the dependence structure from the joint distribution and extricating dependence and marginal behavior leads to one of the major concepts regarding copulas. This important aspect of copulas is based on Sklar's theorem (1959), as follows:

Theorem 1 (Sklar, 1959)**.** *If F is a joint distribution function with marginal distribution functions $F_1, \ldots, F_n$, then there is a copula C such that*

$$F(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n)).$$

The above theorem states that a multivariate distribution can be split into univariate marginal distribution and dependence structure. Specifically, letting f be the probability density function of F, $u_i = F_i(x_i), i = 1, \ldots, n$, and c the density function of copula C,

$$\begin{aligned} f(x_1, \ldots, x_n) &= \frac{\partial F(x_1, \ldots, x_n)}{\partial x_1 \cdots \partial x_n} \\ &= \frac{\partial C(F_1(x_1), \ldots, F_n(x_n))}{\partial c_1 \cdots \partial c_n} \\ &= \frac{\partial C(u_1, \ldots, u_n)}{\partial u_1 \cdots \partial u_n} \times \prod_i \frac{\partial F_i(x_i)}{\partial x_i} \\ &= c(u_1, \ldots, u_n) \times \prod_i f_i(x_i). \end{aligned}$$

Hence, a joint probability density function $f(x_1, \ldots, x_n)$ can be separated into the copula density function and the marginal probability density functions. Here, the correlation structure between variables is specified by the copula density function, c, determining the dependence structure of the variables.

It can easily be shown that $F_i(x_i), i = 1, \ldots, n$, have a uniform distribution defined on $[0, 1]$. Therefore, the copula can be viewed as a multivariate function with a standard uniform marginal distribution. In particular, if the marginal distributions of F are continuous, then F has a unique copula. For such continuous univariate marginal distributions, the unique copula function is given by

$$C(u_1, \ldots, u_n) = F(F_1^{-1}(x_1), \ldots, F_n^{-1}(x_n)),$$

where $F_1^{-1}, \ldots, F_n^{-1}$ denote the quantile functions of the univariate marginal distributions $F_1, \ldots, F_n$. An important feature of a copula is that any choice of marginal distributions can be used. This is due to the fact that a copula links univariate marginal distributions to their multivariate distribution and is independent of marginal distributions. Note that such a copula is constructed provided that marginal distributions are known and can be consistently estimated from data (Nelson, 1999; Embrechts et al., 2003). To summarize, we can choose any copula, leading to different dependence structures once the marginal distributions are determined, but the resulting multivariate distributions have the same marginal distributions. Some examples of copulas considered in this work are presented in the following section.

3.1 Examples

We confine our discussion to the most commonly used family of copulas in modeling multivariate data, the elliptical, to assess dependence structure. Gaussian, Student's t, and Cauchy copulas are all members of the elliptical copula family and will be studied here. The elliptical copulas are known to appropriately validate multivariate dependencies where extreme events often occur. In addition, they are simple in simulation and tail dependence gauge. Before stating the Gaussian and t-copulas, we first describe the simplest copula, the independence copula.

3.1.1 Independence Copula

The Independence copula is simply

$$C(u_1, \ldots, u_n) = u_1 \cdot u_2 \cdots u_n.$$

3.1.2 Gaussian Copula

The copula of the multivariate normal distribution with linear correlation matrix ρ is

$$C_\rho^G(u_1, \ldots, u_n) = G(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n)),$$

where G denotes the joint distribution function of the multivariate standard normal distribution function, and Φ^{-1} is the inverse of the distribution function of the univariate standard normal distribution. A copula of the above form is called Gaussian or Normal copula. When $n = 2$(bivariate distribution case), we obtain the copula function as follows:

$$C_\rho^G(u_1, u_2) = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp\left\{-\frac{s_1^2 - 2\rho s_1 s_2 + s_2^2}{2(1-\rho^2)}\right\} ds_1 ds_2,$$

where the copula parameter (ρ) is estimated by the Spearman's (ρ_s) or Kendall's (ρ_τ) rank correlations calculated from data, as follows:

$$\rho = \arcsin^{-1}(\pi\rho_\tau/2) \quad \text{or} \quad 2\arcsin^{-1}(\pi\rho_s/2).$$

For copulas that have an elliptically symmetric distribution, two coefficients of lower and upper tail dependences of X_1 and X_2, defined in Section 2, are identical, denoted by λ. Note that in the case of a Gaussian copula, the value of λ is 0 (for a proof see Embrechts, McNeil, and Straumann 1999). That is, the Gaussian copula does not have tail dependence, meaning the two random variables are asymptotically independent in the upper and lower tails.

3.1.3 Students' t-copula

Similar to the Gaussian copula, the t-copula is based on the multivariate Student's t distribution. With ν degrees of freedom, a t-copula is defined as

$$C_{\nu,\rho}^t(u_1, \ldots, u_n) = t_{\nu,\rho}^n(t_\nu^{-1}(u_1), \ldots, t_\nu^{-1}(u_n)),$$

Table 1: The coefficients of upper and lower tail dependence of a t-copula for some values of ρ and ν. the last row, with $n \to \infty$, represents the Gaussian copula

$\nu \backslash \rho$	-0.5	0	0.5	0.9	1
1	0.25	0.29	0.5	0.78	1
4	0.01	0.08	0.25	0.63	1
10	0.00	0.01	0.08	0.46	1
∞	0	0	0	0	1

where $t^n_{\nu,\rho}$ is the joint t distribution, and t_ν is the distribution of a standard univariate t distribution. When $n = 2$(the bivariate case), we obtain t-copula as follows:

$$C^t_{\nu,\rho}(u_1, u_2) = \int_{-\infty}^{t_\nu^{-1}(u_1)} \int_{-\infty}^{t_\nu^{-1}(u_2)} \frac{1}{2\pi(1-\rho^2)^{1/2}} \left\{ 1 + \frac{s_1^2 - 2\rho s_1 s_2 + s_2^2}{\nu(1-\rho^2)} \right\}^{-(\nu+2)/2} ds_1 ds_2,$$

where the parameter ν controls the heaviness of the tails.

In contrast with the Gaussian copula which has $\lambda = 0$, the t-copula has a positive value of λ (for a proof see Embrechts, McNeil, and Straumann 1999). Specifically, the coefficient of tail dependence is increasing in ρ and, as one would expect since a t-distribution converges to a normal distribution as ν tends to infinity, decreasing in ν. Hence, we have increasing tail dependence with decreasing parameter ν. Furthermore, the coefficient of upper (lower) tail dependence tends to zero as the number of degrees of freedom tends to infinity for $\rho < 1$. Accordingly, the t-copula has both lower and upper tail dependences. Table 1 shows the coefficient of upper and lower tail dependence, λ, for some values of ρ and λ. Note that the lower the degree of freedom, the heavier the tail dependence is for a t-copula. Therefore, in terms of increasing tail dependence, we expect the rank to be Gaussian, t-copula (ν=4), t-copula (ν=10) and Cauchy copula. In general, as $\nu \to \infty$, $C^t_{\nu,\rho}(u_1, u_2) \to C^G_\rho(u_1, u_2)$. Again, in contrast to the Gaussian copula, a t-copula generates the dependence structure with tail dependence.

3.1.4 Cauchy copula

The Cauchy copula is a special case of the t-copula with $\nu = 1$. Therefore, similar to t-copulas, Cauchy copulas measure the tail dependence structure.

4 Numerical Study

4.1 Simulation

In this section, for simple illustration, we create a bivariate distribution that uses Gaussian and t-copulas, as well as two normal marginal distributions. Note that a bivariate copula is

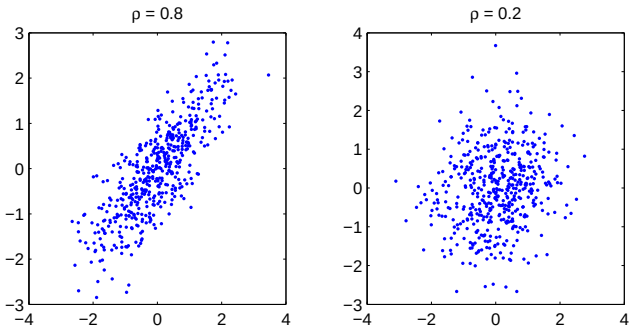


Figure 1: Normal marginal distributions, Gaussian copula, ρ=0.2 and 0.8.

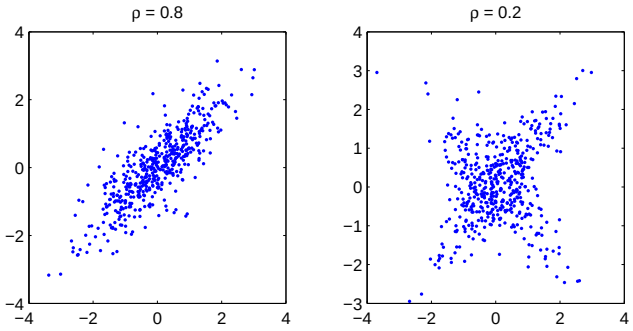


Figure 2: Normal marginal distributions, t-copula with ν=10,ρ=0.2 and 0.8.

a joint probability distribution on two random variables whose distributions are uniform.

A bivariate Gaussian copula is parameterized by the linear correlation matrix, [1 ρ; ρ 1]. Figure 1 shows five hundred simulated points from a Gaussian copula with ρ=0.2 and 0.8, where two standard normal marginal distributions are integrated. Note that any marginal distributions can be used and still have the same correlation. Copulas allow the separation of dependence and marginal distributions. As stated earlier, two random variables approach linear dependence as ρ approaches 1 or -1 and reach independence as ρ approaches zero.

Similar to the Gaussian copula case, five hundred simulated points are produced from a standard bivariate t-copula with three degrees of freedom, ν= 1, 4 and 10, and correlation parameters, ρ = 0.2 and 0.8. That is, three different t-copulas with different degrees of freedom are constructed and still display the same correlation. Figures 2, 3 and 4 shows five hundred simulated values from a bivariate t-copula with the standard normal marginal distributions of the above set up. As in the Gaussian copula, the t-copula has uniform marginal distributions.

The difference between Gaussian copulas and t-copulas is illustrated with the scatter plots above. Figures 1 $\sim$ 4 depict the scatter plots of two random variables with the same

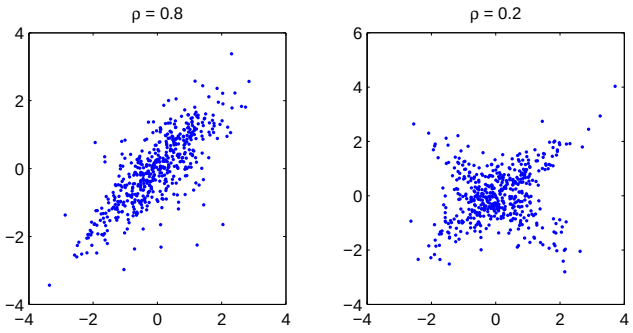


Figure 3: Normal marginal distributions, t-copula with ν=4,ρ=0.2 and 0.8

.

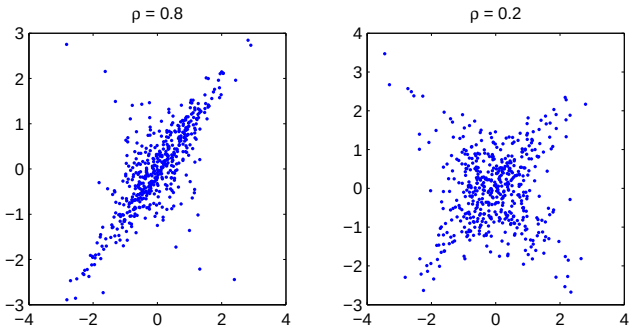


Figure 4: Normal marginal distributions, Cauchy copula, ρ=0.2 and 0.8.

correlation coefficient (0.2 or 0.8). As the scatter plots demonstrate, t-copulas appear to be a bit different, depending on the values of the degree of freedom although their components have the same correlations. Similarly, the t-copulas are unlike the Gaussian copula even when their components have the same correlation. This is due to their differing dependence structures. As mentioned earlier, as the degrees of freedom parameter becomes larger, the t-copula approaches the Gaussian copula.

For a Gaussian copula with normal marginal distributions, the variable distribution is more concentrated. For a Cauchy copula with normal marginal distributions, the variable distribution is more scattered. Therefore, in practice, by choosing an appropriate copula, we can obtain some desired results. For instance, a Cauchy copula would be appropriate in determining economic capital for the portfolio of a multi-line insurance company since it fully captures tail dependence for extreme events (Tang and Valdez, 2006).

4.2 Best Copula

Copulas are in fact the dependence structure of the data distribution that has varying amounts of tail dependence depending on the choice of copulas. Therefore, choosing proper copulas is important, since a poorly chosen copula may lead to decreased sensitivity to the actual association of two variables. The copula selection issue has been studied by many authors, including Genest and Rivest (1993), Frees and Valdez (1998) and Kumar and Shoukri (2008). Here, following Durrleman et al. (2000), as a numerical measure of the quality of fit for data, we employ the discrete L^2 norm evaluating the distance between two copulas. For $X_i^t, i = 1, \ldots, n, t_i = 1, \ldots, T$, the L^2 norm between the estimated copula, C, and the empirical copula, $\hat{C}$, is defined as

$$||C - \hat{C}||_{L^2} = \left\{ \sum_{t_1=1}^{T} \cdots \sum_{t_n=1}^{T} \left[C\left(\frac{t_1}{T}, \ldots, \frac{t_n}{T}\right) - \hat{C}\left(\frac{t_1}{T}, \ldots, \frac{t_n}{T}\right) \right]^2 \right\}^{1/2},$$

where the empirical copula, $\hat{C}$, is given by (Durrleman et al, 2000)

$$\hat{C}\left(\frac{t_1}{T}, \ldots, \frac{t_n}{T}\right) = \frac{1}{T} \sum_{t=1}^{T} \prod_{i=1}^{n} I(r_i^t \leq t_i),$$

where r_i^t, $i = 1, \ldots, n$, is the rank statistic of the sample. This non-parametric version of the copula provides good estimates of the (parametric) copula. Therefore, the best copula is chosen as a minimizer of the above distance, indicating that the chosen copula fits the data best.

4.3 Application

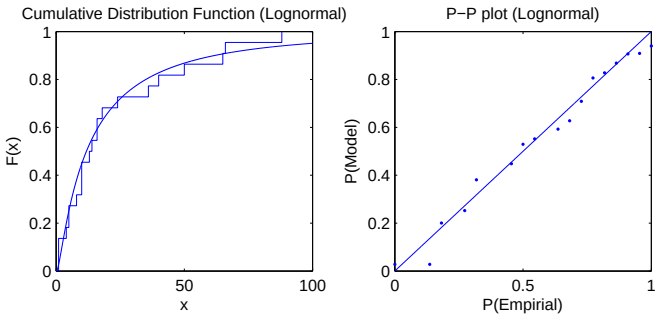


Figure 5: Empirical distribution (left) and P-P plot (right) for multiple myeloma data

In this section, we use the copula model chosen by the L^2 norm criterion to analyze the dependence structure of the multiple myeloma data carried out at the Medical Centre of the University of West Virginia, USA. In this study, survival time of patients (ST), the level of

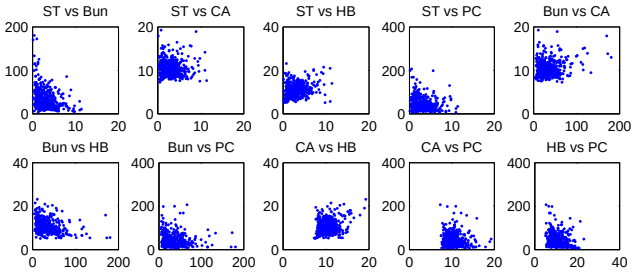


Figure 6: t-copula with ν=4, 500 simulated samples

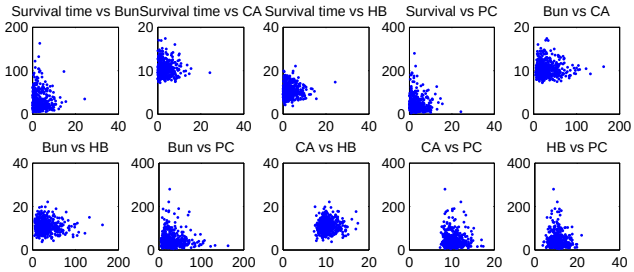


Figure 7: Independence copula, 500 simulated samples

blood urea nitrogen (BUN), serum calcium (CA), hemoglobin (HB), and the percentage of plasma cells (PC) in the bone marrow are considered. The main aim of this data analysis was to investigate the effect of the risk factors BUN, CA, HB and PC on the survival time of patients. Multiple myeloma is a malignant disease caused by the accumulation of abnormal plasma cells (a type of white blood cell) in the bone marrow, resulting in pain and the destruction of bone tissue. For more details, see Collect (1999). Data can be found in Krall et al. (1975). For simplicity, complete data points for males in the data set are used here, where the sample size is 22.

It was found that from the distance measure in Section 4.2, Gaussian, Cauchy, t(ν=4), t(ν=10) and Independence copula yield .9133, .8780, .7168, .8482 and .7363, respectively. This indicates that t-copula with ν=4 may be an appropriate model that fits best for the data. With this copula, some graphical method to further study the dependence of the variables is constructed.

Pearson's correlation coefficients are presented in Table 2, where the numbers in parentheses indicate the p-value under the alternative that two variables are linearly related. As seen from the table, the linear correlation measures are not sufficiently informative on the dependence structure of the variables and are especially problematic to construct the co-movement between extreme (tail) values. For example, the linear correlation coefficient between ST and BUN is -.259 with the p-value of .244, and this does not provide evidence

Table 2: Pearson's correlation coefficient for the multiple myeloma data

	ST	BUN	CA	HB	PC
ST	1	-0.259(0.244)	-0.106(0.640)	0.185(0.409)	-0.233(0.298)
BUN	-0.259(0.244)	1	0.041 (0.857)	-0.235(0.292)	0.287(0.196)
CA	-0.106(0.640)	0.041(0.857)	1	0.165(0.463)	0.148(0.511)
HB	0.185(0.409)	-0.235(0.292)	0.165(0.463)	1	-0.257(0.247)
PC	-0.233(0.298)	0.287(0.196)	0.148(0.511)	-0.257(0.247)	1

that the two variables can be explained via linear dependence structure. As an alternative method to the linear correlation coefficient, we model the dependence structure using copulas.

To begin, we find the appropriate models for the above variables. One well-known graphical method is based on the empirical distribution function of $F(x)$, defined as

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

for $X_1, \ldots, X_n$. the plot of $\hat{F}$ versus F should be close to each other if the assumed model is legitimate. For example, Figure 5 (left) shows that the empirical distribution for the survival time is fairly close to the specified distribution, the lognormal distribution. Hence the lognormal distribution appears to be an appropriate distributional model. Another method is the probability-probability (P-P) plot which will be approximately linear if the assumed model is correct. Figure 5 (right) shows that the reference diagonal line, along with the graph points should fall for the survival time. There appears no significant linear departure from the straight line. Thus the lognormal distribution for the survival time seems reasonable, which is in agreement with the previous result. In addition, we performed some famous numerical model checking methods such as the Anderson-Darling test, Kolmogorov-Smirnov test and Chi-Square test. Judging from these, along with the graphical methods, we arrive at the distributional results summarized in Table 3.

Associated with the marginal distributions above, we generated 500 simulated samples under t-copula with ν=4. It is important to note that even with identical correlations, the dependence structure may be different depending on the choice of copula, especially on tail behaviors, leading to different risk levels. Specifically, t-copula captures non-linear trends well, while Gaussian copula fits well only when dependence is mostly linear.

Although, as the degrees of freedom decrease, the tail dependence becomes more remarkable, care must be taken when using t-copula. This is because poorly chosen degrees freedom may lead to some undesirable results due to a crossing distribution, where the positive and negative values in the sample are not distinguishable. In Figure 6 with t-copula

Table 3: Distribution and estimate for the multiple myeloma data

	Distribution	Estimate	Mean	Std. Dev.	Skewness	Kurtosis
ST	Lognormal	μ=2.47, σ=1.29	22.77	24.42	1.44	1.26
BUN	Lognormal	μ=3.28, σ=0.66	34.0	32.44	3.37	13.50
CA	Gumbel (Max)	μ=9.59, σ=1.26	10.32	1.62	1.37	2.14
HB	Gamma	α=14.79, β=0.73	10.43	2.77	-0.14	-0.89
PC	Lognormal	μ=3.55, σ=0.65	42.59	27.45	0.94	0.01

(ν=4), it seems that the positive effect is somewhat stronger than the negative effect in the scatter plots of ST vs CA, ST vs HB, ST vs PC, CA vs HB, and CA vs PC. In particular, a large amount of serum calcium (CA) seems to significantly influence the survival time (ST) of patients. As expected, the independence copula (Figure 7) provides no distinct patterns due to the assumption of independence among variables.

5 Conclusion

Copulas provide a useful way to model the joint distribution of two or more random variables. We have discussed the advantages of using copulas in statistical modeling, summarized as follows: both linear and non-linear dependence structure can be detected, arbitrary choice of a marginal distribution is allowed, and tail dependence between variables can be explored with more flexibility and efficiency. In applying copulas to a real world data set regarding multiple myeloma, we have observed that the qualitative difference between the variables can be detected using the t-copula with degree of freedom equal to 4, whereas the Gaussian fails to capture this. The results show that tail dependence is a feature of the data set, so the t- copula is a suitable choice.

Acknowledgements
The authors are grateful to the referees and the editor for their helpful comments and reviews of the manuscript.

References

[1] Breymann, W., Dias, A., and Embrechts, P. (2003). Dependence Structures for Multivariate High-Frequency Data in Finance. *Quantitative Finance*, **3**, 1–16.

[2] Collect, D. (1999). *Modelling Survival Data in Medical Research*. Chapman.

[3] Durrleman, V., Nikeghbail, A., and Roncalli, T. (2000). Which copula is the right one?. Credit Lyonnais, Available at SSRN: http://ssm.com/abstract=1032545.

[4] Embrechts, p., McNeil, A., and Straumann, D. (1999). *Correlation and Dependence in Risk Management: Properties and Pitfall. Risk Management: Value at Risk and Beyond*. Cambridge Publication.

[5] Embrechts, P., Lindskog, F., and McNeil, A. (2003). *Modelling Dependence with Copulas and Applications to Risk Management*. Handbook of Heavy Tailed Distributions in Finance, ed. S. Rachev, Elsevier, Chapter 8, 329–384.

[6] Frees, E. and Valdez, E. (1998). Understanding Relationship Using Copulas. *North American Actuarial Journal*, **2**, 1-25.

[7] Genest, C. and Rivest, L. (1993). Statistical inference procedures for bivariate Archimedean copulas. *Journal of American Statistical Association*, **88**, 1034–1043.

[8] Krall, J. M., Uthoff, V. A., and Harley, J. B. (1975). A step-up procedure for selecting variables associated with survival. *Biometrics*, **31**, 49–51.

[9] Kumar, P. and Shoukri, M. (2007). Evaluating Aortic Stenosis Using the Archimedean Copula Methodology. *Journal of Data Science*, **6**, 173–187.

[10] Lehmann, E. (1966). Some concepts of dependence. *Annals of Mathematical Statistics*, **37**, 1137–1153.

[11] Nelsen, R. (1999). *An Introduction to Copulas*. Springer Verlag, New York.

[12] Schweizer, B. and Wolff, E. F. (1981). On the Nonparametric Measures of Dependence for Random Variables. *Annals of Statistics*, **9**, 879–885.

[13] Schweizer, B. (1991). Thirty Years of Copulas. Advances in Probability Distributions with Given Marginals (G. Dall'Aglio, S.Kotz and G. Salinetti, eds.), 13–50, Kluwer, Dordrecht.

[14] Sklar, A. (1959). Functions de Repartition a n Dimensions et leurs Merges. Publication of *the Institute of Statistics, University of Paris 8*, 229–231.

[15] Shim, J.B., Lee, S. and MacMinn, R. (2009). Measuring Economic Capital: Value at Risk, Expected Tail Loss and Copula Approach. Working paper.

[16] Tang, A. and Valdez, E. A. (2006). Economic Capital and the Aggregation of Risks Using Copulas. 28th International Congress of Actuaries. Working paper.

[17] Zheng, M. and Klein, J. (1995). Estimates of Marginal Survival for Dependent Competing Risks Based on an Assumed Copula. *Biometrika*, **82**, 127–138.

Journal of Statistical Research *ISSN 0256 - 422 X*
2009, Vol. 43, No. 2, pp. 17-27
Bangladesh

BAYESIAN MULTIVARIATE LATENT VARIABLE MODEL FOR MIXED CORRELATED ORDINAL AND CONTINUOUS RESPONSES

E. Bahrami Samani
Department of Statistics, Faculty of Mathematical Science
Shahid Beheshti University, Tehran, 1983963113, Iran
Email: ehsan_bahrami_samani@yahoo.com

M. Ganjali
Department of Statistics, Faculty of Mathematical Science
Shahid Beheshti University, Tehran, 1983963113, Iran
Email: m-ganjali@sbu.ac.ir

SUMMARY

A general framework is proposed for joint modelling mixed correlated ordinal and continuous responses. A Markov Chain Monte Carlo sampling algorithm is described for estimating the posterior distribution of the parameters. Because of the flexibility of the modelling framework and estimation procedure, extensions to ordered categorical outcomes and more complex data structures are straightforward. The methods are illustrated by using a large data set excerpted from the British Household Panel Survey (BHPS). For these data, the simultaneous effects of some covariates on life satisfaction, income and the amount of money spent on leisure activities per month as three mixed correlated responses are explored.

Keywords and phrases: Joint modelling, Gibbs sampler, Multivariate latent variable, Ordinal and continuous responses, Markov Chain Monte Carlo.

1 Introduction

Rectangle probabilities from the multivariate normal distribution have many applications in statistics. This include the multivariate latent variable model for mixed correlated ordinal and continuous responses. In general this probabilities require multidimensional integrals, which can be evaluated by multidimensional numerical methods. The use of numerical methods for parameters estimation was too time consuming for dimensions greater than five. Numerical methods has computational time (and memory requirements) exponentially increasing in the dimension of the integrals. So, outcomes measured on a variety of scales (mixed continuous and ordinal), can be difficult for joint modelling of responses.

For joint modelling of responses, one method is to use the general location model of Olkin and Tate (1961), where the joint distribution of the continuous and categorical variables is decomposed into a marginal multinomial distribution for the categorical variables and a conditional multivariate normal distribution for the continuous variables, given the categorical variables (for a mixed poisson and continuous responses where Olkin and Tate's method is used see Yang et al., 2007 and for joint modelling of mixed outcomes using latent variables see McCulloch, 2007). A second method for joint modelling is to decompose the joint distribution as a multivariate marginal distribution for the continuous responses and a conditional distribution for categorical variables given the continuous variables. Cox and Wermuth (1992) empirically examined the choice between these two methods. The third method uses simultaneous modelling of categorical and continuous variables to take into account the association between the responses by the correlation between errors in the model for responses. For more details of this approach see, for example, Heckman (1978) in which a general model for simultaneously analyzing two mixed correlated responses is introduced and Catalano and Ryan (1992) who extended and used the model for a cluster of discrete and continuous outcomes.

In this paper, a new class of latent variable models is proposed for mixed correlated ordinal and continuous responses, and Markov chain Monte Carlo (MCMC) algorithms (Tierney, 1994) are developed for estimating the posterior distribution of the parameters. The aim of this paper is to use and extend an approach similar to that of Heckman (1978), which jointly models a nominal and a continuous variable, for joint modelling of multivariate ordinal and continuous outcomes. However, the Bayesian approach has several important advantages. First, the exact posterior distribution of the parameters can be estimated by using MCMC methods. Means and quantities based on the estimated posterior are appropriate regardless of the sample size. In contrast, standard errors and confidence limits for the maximum likelihood estimates are typically based on strong asymptotic normality assumptions. Second, the Bayesian approach allows for the direct incorporation of prior knowledge. This is a major advantage in structural equation modelling. Classical methods often require that a subset of the parameters is known to ensure identifiability. Although constrain on the threshold parameters and the variance of the latent variables are often reasonable, additional less justifiable constraints can be avoided by using a prior distibution to allow for prior uncertainty in the parameters. In addition, by assigning an informative prior to parameters about which there is previous informative, more precise estimates of the parameters of interst can be obtained.

In Section 2, the general modelling framework is described. A general MCMC sampling algorithm for posterior estimation is outlined in Section 3. In Section 4, we have a simulation study. In Section 5, the proposed methodology is applied on the BHPS data. Finally, concluding remarks are given.

2 Latent Variable Model

We use Y_{ij} to denote jth ordinal response for the ith individual with c_j levels defined as,

$$Y_{ij} = \begin{cases} 1, & \text{if } Y_{ij}^* < \theta_{1,j}, \\ k+1, & \text{if } \theta_{k,j} \leq Y_{ij}^* < \theta_{k+1,j}, \quad k = 1, \ldots, c_j - 2 \\ c_j, & \text{if } Y_{ij}^* \geq \theta_{c_j,j}, \end{cases}$$

where $\theta_{1,j}, \ldots, \theta_{c_j-1,j}$ are the cut-point parameters and Y_{ij}^* denotes the underlying latent variable for Y_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, M_1$. The joint model takes the form:

$$\left.\begin{array}{l} Y_{ij}^* = \beta_j' X_i + \varepsilon_{ij}^{(1)}, \quad j = 1, \ldots, M_1 \\ Z_{ij} = \beta_j' X_i + \varepsilon_{ij}^{(2)}, \quad j = M_1 + 1, \ldots, M \\ \varepsilon_i = (\varepsilon_i^{(1)}, \varepsilon_i^{(2)})' \overset{iid}{\sim} MVN(0, \Sigma) \end{array}\right\} \tag{2.1}$$

where $\varepsilon_i^{(1)} = (\varepsilon_{i1}^{(1)}, \ldots, \varepsilon_{iM_1}^{(1)})', \varepsilon_i^{(2)} = (\varepsilon_{i(M_1+1)}^{(2)}, \ldots, \varepsilon_{iM}^{(2)})'$, $\theta_j = (\theta_{1,j}, \ldots, \theta_{c_j-1,j})', j = 1, \ldots, M_1$, is the vector of cutpoint parameters for the jth ordinal response and X_i is the vector of explanatory variables for the i^{th} individual and Σ is th $M \times M$ covariance matrix which for illustration, when $M_1 = 2$ and $M = 3$ has the following structure,

$$\Sigma = \begin{pmatrix} 1 & \rho_{12} & \sigma\rho_{13} \\ \rho_{12} & 1 & \sigma\rho_{23} \\ \sigma\rho_{13} & \sigma\rho_{23} & \sigma^2 \end{pmatrix},$$

where σ^2 is the variance of the continuous response, and $\rho_{jj'}$ for $j \neq j'$, $j = j' = 1, 2, 3$ is the correlation between jth and j'th responses. The vector of coefficients β_j, cutpoints parameters θ_j for $j = 1, \ldots, M_1$ and Σ should be estimated. The parameter vector, β_j for $j = M_1 + 1, \ldots, M$, includes an intercept parameter but β_j, for $j = 1, \ldots, M_1$, due to having cutpoints parameters, are assumed not to include any intercept. In this model any multivariate distribution can be assumed for the errors in the model. Here, a multivariate normal distribution is used.

3 Bayesian Estimation

In this section, prior distributions are chosen for the parameters, and general **MCMC** algorithm is outlined for estimating the posterior distributions of the parameters and the latent variables (via, Dunson, 2000). The prior distributions are conjugate if the underlying variables are normal.

Markov chain Monte Carlo (**MCMC**) methods use computer simulation of Markov chains in the parameter space. The Markov chains are defined in such a way that the

posterior distribution, in given statistical inference problem, is the asymptotic distribution. One of the standard approach to define such Markov chains is Gibss sampling. We will use **MCMC** techniques for posterior computation in the models proposed in section 2. In the special case where all the underlying and latent variables have normal distribution the **MCMC** algorithm is Gibbs sampler that follows a simple from.

3.1 Prior Distributions

The parameters $\eta = (\beta_1, \ldots, \beta_M, \theta_1, \ldots, \theta_{M_1})'$ are assigned a normal prior $\eta \sim N_q(\mu_0, \Sigma_0)$, where μ_0 is a vector of location parameters, Σ_0 is a covariance matrix and $q = M + M_1$. To choose a vague prior distribution for θ, set $\mu_0 = 0$ and $\Sigma_0 = \text{diag}\{\sigma_1^2, \ldots, \sigma_q^2\}$. Wishart prior are specified for the the precision matrix Σ in expressions (1):

$$\Sigma \sim \text{Wishart}(\nu, \Lambda)$$

with degrees of freedom ν and precision Λ . A prior can be assigned by choosing $\nu \geq M$, where M is the the dimension of Σ.

3.2 The Gibbs Sampler

Consider model with all equations given in (2.1). We assume that η and Σ are a priori independent with $p(\eta) = N_q(\mu_0, \Sigma_0)$ and $p(\Sigma) = \text{Wishart}(\nu, \Lambda)$, where $p(.)$ is the prior distribution. The conditional posterior distribution of $p(\eta|y, z, \Sigma)$ and $p(\Sigma|y, z, \eta)$ are computed in section 3.3.

To estimate the posterior distributions of the parameters, we define a Morkov chain in $\xi = (\eta, \Sigma)$. Denote with $\xi^{(t)} = (\eta^{(t)}, \Sigma^{(t)})$ the state parameter of the Markov chain after t iterations. Given the nature of a Morkov chain, all we need to define is the transition probability, i.e., given a current value for $\xi^{(t)}$, we need to generate a new value $\xi^{(t+1)}$. We do so by sampling from the complete conditional posterior distributions for η and Σ

$$\eta^{(t+1)} \sim p(\eta^{(t)}|y, z, \Sigma^{(t)})$$
$$\Sigma^{(t+1)} \sim p(\Sigma^{(t)}|y, z, \eta^{(t)}).$$

Step 1 through 2 define a Morkov chain $\xi^{(t)}$ which converges to $p(\eta, \Sigma|y, z)$, as desired. The described Markov chain Monte Carlo simulation is a special case of a Gibbs sampler. In general, let $\xi^* = (\xi_1, \ldots, \xi_p)$ denote the parameter vector. The Gibbs sampler proceeds by iteratively, for $j = 1, \ldots, p$, generating from the conditional posterior distributions

$$\xi_j^{(t+1)} \sim p(\xi|\xi_1^{(t+1)}, \ldots, \xi_{j-1}^{(t+1)}, \xi_{j+1}^{(t)}, \ldots, \xi_p^{(t)}, y, z).$$

3.3 Posterior Computations

In this section, an **MCMC** algorithm is outlined for posterior computation of model given in equations (2.1). We apply the following definition and theorem for multivariate distributions to obtain the form of the joint posterior distribution.

Definition 3.1. If $F(w_1, \cdots, w_{M_1}) = P(W_1^* \leq w_1, \ldots, W_{M_1}^* \leq w_{M_1})$ is a distribution function, operator $\Delta_{b_j a_j} F(w_1, \ldots, w_{M_1})$ is defined as, $(a_j \leq b_j)$

$$F(w_1, \ldots, w_{(j-1)}, b_j, w_{(j+1)}, \ldots, w_{M_1}) - F(w_1, \ldots, w_{(j-1)}, a_j, w_{(j+1)}, \ldots, w_{M_1}).$$

Theorem 3.1. If for $j = 1, \ldots, M_1$, $a_j \leq b_j$, then

$$P(a_1 < W_1^* \leq b_1, \ldots, a_{M_1} < W_{M_1}^* \leq b_{M_1}) = \Delta_{b_1 a_1} \cdots \Delta_{b_{M_1} a_{M_1}} F(w_1, \cdots, w_{M_1}),$$

where $\Delta_{b_1 a_1} \cdots \Delta_{b_{M_1} a_{M_1}} F(w_1, \cdots, w_{M_1}) = F_0 - F_1 + F_2 - \cdots + (-1)^{M_1} F_{M_1}$; and F_j is the sum of all $\binom{M_1}{j}$ terms of the from $F(g_1, \ldots, g_{M_1})$ with $g_k = a_k$ for exactly j integers in $\{1, \ldots, M_1\}$, and $g_k = b_k$ for the remaining $M_1 - j$ integers.

Proof. See Ash and Dolens-Dade (2000, page 27). □

Let $y = (y_1', \ldots, y_n')'$, $z = (z_1', \ldots, z_n')'$ and $x = (x_1', \ldots, x_n')'$ where $y_i = (y_{i1}, \ldots, y_{iM_1})'$, $z_i = (z_{i(M_1+1)}, \ldots, z_{iM})'$ and $x_i = (x_{i1}, \ldots, x_{ip})'$, and p is the number of explanatory variables for the i^{th} individual (the number of components in this vector may also be dependent on the chosen variable, i.e. x_i be x_{im} and p be p_m, here, we ignore this for simplicity).

The joint posterior distribution for the parameters and latent variables is:

$$\begin{aligned}
P(\eta, \Sigma \,|y, z\,, x) &\propto f_{y,z}(y, z|\eta, \Sigma, x).\pi(\eta, \Sigma|x) \\
&\propto \left[\prod_{i=1}^{n} f(z_i,\ y_i|x_i, \eta, \Sigma)\right]\pi(\eta, \Sigma|x) \\
&\propto \left[\prod_{i=1}^{n} P(Y_{i1} = y_{i1}, \ldots, Y_{iM_1} = y_{iM_1}|z_i, x_i) f(z_i|x_i)\right]\pi(\eta, \Sigma|x) \\
&\propto \left[\prod_{i=1}^{n} P(\theta_{1,y_{i1-1}} < Y_{i1}^* \leq \theta_{1,y_{i1}}, \ldots, \theta_{M_1,y_{i,M_1-1}} < Y_{iM_1}^* \leq \theta_{M_1,y_{iM_1}}|z_i, x_i)\right. \\
&\qquad \left. f(z_i|x_i)\right]\pi(\eta, \Sigma|x)
\end{aligned}$$

where $\pi(.|x)$ denote the joint prior density and $\eta = (\beta_1, \ldots, \beta_M, \theta_1, \ldots, \theta_{M_1})'$.

Using the above Theorem, the joint posterior distribution could be summarized as,

$$\begin{aligned}
P(\eta, \Sigma \,|y, z\,, x) &\propto \left[\prod_{i=1}^{n} \Delta_{b_{i1} a_{i1}} \cdots \Delta_{b_{iM_1} a_{iM_1}} F(w_{i1}, \cdots, w_{iM_1}|z_i, x_i) f(z_i|x_i)\right]\pi(\eta, \Sigma|x) \\
&\propto \left[\prod_{i=1}^{n} (F_{i0} - F_{i1} + F_{i2} - \cdots + (-1)^{M_1} F_{iM_1})\right] |\Sigma|^{(\nu - M - 1)/2} |\Sigma_{22}|^{-n(M - M_1)/2} \\
&\qquad \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (z_i - \mu_{z_i})' \Sigma_{22}^{-1} (z_i - \mu_{z_i}) - \frac{1}{2} (\eta - \mu_0)' \Sigma_0^{-1} (\eta - \mu_0) \right\},
\end{aligned}$$

where $\mu_{z_i} = (\beta'_{M_1+1}X_i, \ldots, \beta'_{M_1+1}X_i)'$, $\Sigma_{22} = Var(Z_i)$, $b_{ij} = \theta_{j,y_{ij}}$ and $a_{ij} = \theta_{j,y_{ij}-1}$ and F_{ij} is the sum of all $\binom{M_1}{j}$ terms of the from $F(g_{i1}, \ldots, g_{iM_1}|z_i, x_i)$ with $g_{ik} = a_{ik}$ for exactly j integers in $\{0, 1, \ldots, M_1\}$, and $g_{ik} = b_{ik}$ for the remaining $M_1 - j$ integers.

We require the full conditional distributions of each parameters of the unknowns. We have

$$P(\eta\,|y, z, \Sigma\,, x) \propto \Big[\prod_{i=1}^{n}(F_{i0} - F_{i1} + F_{i2} - \cdots + (-1)^{M_1}F_{iM_1})\Big]\,|\Sigma_{22}|^{-n(M-M_1)/2}$$
$$\exp\Big\{-\frac{1}{2}\sum_{i=1}^{n}(z_i - \mu_{z_i})'\Sigma_{22}^{-1}(z_i - \mu_{z_i}) - \frac{1}{2}(\eta - \mu_0)'\Sigma_0^{-1}(\eta - \mu_0)\Big\}.$$

The full conditional distribution of the precision matrix Σ is

$$P(\Sigma\,|y, z, \eta\,, x) \propto \Big[\prod_{i=1}^{n}(F_{i0} - F_{i1} + F_{i2} - \cdots + (-1)^{M_1}F_{iM_1})\Big]\,|\Sigma_{22}|^{-n(M-M_1)/2}$$
$$\exp\Big\{-\frac{1}{2}\sum_{i=1}^{n}(z_i - \mu_{z_i})'\Sigma_{22}^{-1}(z_i - \mu_{z_i})\Big\}|\Sigma|^{(\nu-M-1)/2}.$$

4 Simulation

We consider three continuous variables Y_1^*, Y_2^* and Z. The ordinal variable Y_1 and Y_2 with three levels is defined as

$$Y_1 = \begin{cases} 1, & \text{if } Y^* < \theta_1, \\ 2, & \text{if } \theta_1 \le Y^* < \theta_2, \\ 3, & \text{if } Y^* \ge \theta_2, \end{cases} \text{ and } Y_2 = \begin{cases} 1, & \text{if } Y^* < \eta_1, \\ 2, & \text{if } \eta_1 \le Y^* < \eta_2, \\ 3, & \text{if } Y^* \ge \eta_2, \end{cases}$$

the variables Z, Y_1^* and Y_2^* are generated by a multivariate normal distribution with zero mean and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}.$$

We consider two sets of cut points. In the first set we let $\theta_1 = -1$ and $\theta_2 = 1$. In the second set we let $\eta_1 = -1$ and $\eta_2 = 1$. In the first and second set of cut points, not having any covariate in the model for latent variable of Y_1 and Y_2, one expect to have, roughly, 16 percents of Y values to be equal to 1, 16 percent to be equal to 3 and 68 percent to be equal to 2. So, the low and high values have nearly the same frequency but the middle value have the highest frequency. For this we consider 3 values for n (50, 100, and 1000). In this

analysis we use 1000 sets of simulation. In each simulation we analyze the following simple model

$$Y_1^* = \varepsilon_1, \quad Y_2^* = \varepsilon_2, \quad Z = \mu_z + \varepsilon_3.$$

Table 1 contains the average estimated values of μ_z, σ_z^2, ρ_{12} (the correlation between Z and Y_1^*), ρ_{13} (the correlation between Z and Y_2^*), ρ_{23} (the correlation between Y_1^* and Y_2^*), θ_1, θ_2, η_1 and η_2 for n=50, n=100 and n=1000. The parameter estimates by the model for μ_z, σ_z^2, ρ_{12}, ρ_{13}, ρ_{23}, θ_1, θ_2, η_1 and η_2 (for n = 50, n=100 and n=1000) are close to the true values of the parameters. Of course, the more the value of n the better the estimates. We used a Gibbs sampler within **winBUGS** to estimate from the joint posterior distribution of the parameters. We run three chains with widely varying intial values and used 10000 Gibbs iterates collected after convergence from each chain to compute posterior summaries of the parameters. posterior summaries of the global parameters for each outcome are shown in Table I.

Table 1: Results of the simulation study

		n=50		n=100		n=1000	
Parameter	True value	Est.	S.D	Est.	S.D.	Est.	S.D.
μ_z	0.000	0.047	0.140	0.010	0.100	0.001	0.023
σ_z^2	1.000	1.121	0.105	1.012	0.072	1.001	0.025
ρ_{12}	0.500	0.479	0.139	0.495	0.081	0.501	0.028
ρ_{13}	0.500	0.488	0.126	0.490	0.091	0.505	0.017
ρ_{23}	0.500	0.469	0.148	0.492	0.071	0.509	0.011
θ_1	-1.000	-1.142	0.222	-0.983	0.152	-0.995	0.045
θ_2	1.000	1.075	0.282	1.020	0.163	0.996	0.030
η_1	-1.000	-1.172	0.233	-0.963	0.140	-0.997	0.048
η_2	1.000	1.091	0.225	1.029	0.144	0.986	0.057

5 Application

5.1 Data

The data used in this paper is excerpted from the 15th wave (2005) of the British Household Panel Survey (BHPS); a longitudinal survey of adult Britons, being carried out annually since 1991 by the ESRC UK Longitudinal Studies Center with the Institute for Social and Economical Research at the University of Essex. These data are recorded for 11251 individuals. The selected variables which will be used in this application are exp1aind in

the following. One of the responses is the life Satisfaction (LS), [where the related question is QA: "How dissatisfied or satisfied are you with your life overall?"] which is measured by directly asking the level of an individual's satisfaction with life overall, resulting in a three categories ordinal variable [1: Not satisfied at all (10.300%). 2: Not satis/dissat (45.400%) and 3: Completely satisfied (44.300 %)]. In our application, the percentage of missing values of LS is 2.000%, so with ignoring the missing values only complete eases are used in our analysis. The amount of money spent on leisure activities per month including money spent on entertainment and hobbies (AM) is also measured [where the related question is QB: "Please look at this Response categories/range and tell me about how much you personally spend in an average month on leisure activities, and entertainment and hobbies, other than eating out?"] as an ordinal response with three categories, [0: Nothing (17.515 %). 1: Under 50 Pound (53.449%) and 2: 50 Pound or over (29.036%.)]. Moreover, the exact amount of an individuals annual income (INC) in the past year in thousand pounds, considered here in the logarithmic scale, is also excerpted as a continuous response variable (mean: 4.068). These three responses, LS, AM and logarithm of income are endogenous correlated variables and should be modelled as a multivariate vector of responses. Socio-demographic characteristics, namely: Gender (male: 44.200% and female: 55.800%), Marital Status (MS)[married or living as couple: 68.500%, widowed: 8.300%, divorced or separated: 8.400% and never married: 14.800%], Age (mean: 49.180) and Highest Educational Qualification (HEQ)[higher or first degree: 15.100%, other higher QF: 64.600%, other QF: 2.000% and no qualification: 18.300%] are also included in the model as covaiates. The vector of explanatory variables is $X = (Gender, Age, MS_1, MS_2, MS_3, HEQ_1, HEQ_2, HEQ_3)$, where MS_1, MS_2 and MS_3 are dummy variables for married or living as couple, widowed and divorced or separated, respectively, and HEQ_1, HEQ_2, HEQ_3 are dummy variables for higher or first degree, other higher QF and other QF, respectively.

5.2 Models for BHPS Data

We apply the model described in section 2 to evaluate the effect of Age, Gender, HEQ and MS simultaneously on LS, AM and Income. We shall also try to find answers for some questions, including (1) do male's LS, AM and income differ from female's? (2) How does HEQ affect the three responses? (3) do a significant correlations exist between three responses? (4) what would be the consequence of not considering these correlations?

The model, which takes into account the correlations between three errors, is:

$$\begin{aligned} LS^* &= \beta_{11}\, MS_1 + \beta_{12}\, MS_2 + \beta_{13}\, MS_3 + \beta_{14}\, HEQ_1 + \beta_{15}\, HEQ_2 + \beta_{16}\, HEQ_3 \\ &\quad + \beta_{17} Gender + \beta_{18} AGE + \varepsilon_1 \\ AM^* &= \beta_{21}\, MS_1 + \beta_{22}\, MS_2 + \beta_{23}\, MS_3 + \beta_{24}\, HEQ_1 + \beta_{25}\, HEQ_2 + \beta_{26}\, HEQ_3 \\ &\quad + \beta_{27} Gender + \beta_{28} AGE + \varepsilon_2 \\ \log(INC) &= \beta_{30} + \beta_{31}\, MS_1 + \beta_{32}\, MS_2 + \beta_{33}\, MS_3 + \beta_{34}\, HEQ_1 + \beta_{35}\, HEQ_2 \\ &\quad + \beta_{36}\, HEQ_3 + \beta_{37} Gender + \beta_{38} AGE + \varepsilon_3. \end{aligned}$$

In the third above equation the logarithm is taken in the base e. For this model the covariance matrix takes the form,

$$\Sigma = \begin{pmatrix} 1 & \rho_{12} & \sigma\rho_{13} \\ \rho_{21} & 1 & \sigma\rho_{23} \\ \sigma\rho_{31} & \sigma\rho_{32} & \sigma^2 \end{pmatrix}.$$

Here, a multivariate normal distribution with correlation parameters ρ_{12}, ρ_{13} and ρ_{23} is assumed for the errors and these parameters should be also estimated.

5.3 Results

We used a Gibbs sampler within **winBUGS** to estimate from the joint posterior distribution of the parameters. We run three chains with widely varying intial values and used 10000 Gibbs iterates collected after convergence from each chain to compute posterior summaries of the parameters. posterior summaries of the global parameters for each outcome are shown in Table 2.

As it can be seen, two correlation parameters ρ_{12} and ρ_{23} are strongly significant. They show a positive correlation between LS and AM ($\rho_{12} = 0.136$, P-value = 0.000) and a positive correlation between log(INC) and AM ($\rho_{23} = 0.136$, P-value = 0.000). Model shows a significant effect of age (the older the individual the more the life satisfaction), MS (married people are more satisfied than never married people and divorced or separated people are less satisfied than never married people), HEQ (the higher the qualification the higher the life satisfaction) and gender (males are more satisfied than females) on the life satisfaction status. All explanatory variables have significant effect on the ordinal response of amount of money spent on leisure activities. Never married people spend more on leisure time activities than other people. The higher the education the more the leisure time activities. Females spend more amount of money than males for leisure time and older people spend less money than younger ones. Also the effect of all explanatory variables are significant on the logarithm of income. Parameter estimates indicate that as the degree of educational qualification increases log(INC) increases. Never married people have less logarithm of income than married people and divorced or separated people. Females have more logarithm of income than males and the older people earn less money than younger ones.

6 Conclusion

In this paper Bayesian multivariate latent variable model is presented for simultaneously modelling of ordinal and continuous correlated responses. We assume a multivariate normal distribution for errors in the model. However, any other multivariate distribution such as t or logistic can be also used. Binary responses are a special case of ordinal responses. So, our model can also be used for mixed binary and continuous responses (Cox and Wermuth,

1992). For correlated nominal, ordinal and continuous responses DeLeon and Carriégre (2007) have developed a model by extending general location model. Generalization of our model for nominal, ordinal and continuous responses is an ongoing research on our part.

References

[1] Ash, R. B. and Dolens-Dade, C. A. (2000). *Probability and measure thoery.* Academic Press.

[2] Catalano, P. and Ryan, L. M. (1992). Bivariate latent variable models for clustered discrete and continuous outcoms. *Journal of the American Statistical Association*, **50**, 3, 1078–1095.

[3] Cox, D. R, and Wermuth, N. (1992). Response models for mixed binary and quantitative variable. *Biometrika*, **79**(3), 441–461.

[4] DeLeon, A. R. and Carriégre, K. C. (2007). General mixed-data model: extension of general location and group continuous models. *Canadian Journal of Statistics*, **35**(4), 533-548.

[5] Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society, Series B*, **62**, 355–366.

[6] Heckman, J. J. (1978). Dummy endogenous variable in a simutaneous equation system. *Econometrica*, **46**(6), 931–959.

[7] McCulloch, C. (2007). Joint modelling of mixed outcome type using latent variables. *Statistical Methods in Medical Research* , **17**(1), 53–74.

[8] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **82**, 669–710.

[9] Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Mathematical Statistics*, **22**, 1701–1786.

[10] Olkin, L. and Tate R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics*, **32**, 448–456.

[11] Yang, Y., Kang, J., Mao, K. and Zhang, J. (2007). Regression models for mixed Poisson and continuous longitudinal data. *Statistics in Medicine*, **26**, 3782–3800.

Table 2: Posterior summaries of the parameters for BHPS data [Marital status (baseline: Never married), Highest education qualification (baseline: No QF), Gender (baseline: Female); parameter estimates highlighted in **bold** are significant at 5 % level.]

Variables	Mean	S.D.
Response: LS		
Married or Living as Couple	**0.202**	0.029
Widowed	0.031	0.039
Divorced or Separated	**-0.359**	0.044
Higher or First Degree	**-0.063**	0.009
Other higher QF	**-0.132**	0.006
Other QF	**-0.209**	0.100
Male	**0.039**	0.021
AGE	**0.005**	0.001
cutpoint 1	**-0.988**	0.102
cutpoint 2	**0.449**	0.119
Response: AM		
Married or Living as Couple	**-0.187**	0.032
Widowed	**-0.238**	0.055
Divorced or Separated	**-0.258**	0.047
Higher or First Degree	**0.578**	0.095
Other higher QF	**0.287**	0.091
Other QF	**-0.078**	0.008
Male	**-0.450**	0.022
AGE	**-0.014**	0.001
cutpoint 1	**-2.333**	0.105
cutpoint 2	**-0.692**	0.122
Response: log(INC)		
Constant	**4.245**	0.039
Married or Living as Couple	**0.114**	0.011
Widowed	**0.217**	0.019
Divorced or Separated	**0.215**	0.016
Higher or First Degree	**0.391**	0.036
Other higher QF	**0.177**	0.035
Other QF	0.031	0.035
Male	**-0.227**	0.007
AGE	**-0.002**	0.001
Variance of log(INC)	**0.153**	0.002
Corr(LS*,AM*)	**0.136**	0.012
Corr (LS *,INC)	0.001	0.001
Corr(AM *,INC)	**0.138**	0.010

Journal of Statistical Research *ISSN 0256 - 422 X*
2009, Vol. 43, No. 2, pp. 29-35
Bangladesh

ASYMPTOTIC DISTRIBUTION OF THE NUMBER OF RECTANGLES ARISING IN BINOMIAL TRIALS

Zafar Iqbal
National College of Business Administration and Economics, Lahore, Pakistan
Email: zafariqbal75@yahoo.com

Ahmed Zogo Memon
National College of Business Administration and Economics, Lahore, Pakistan
Email: azogomemon@hotmail.com

SUMMARY

The simultaneous occurrence of independent and identical Binomial trials at n^2 locations of an $n \times n$ lattice may randomly produce a specified configuration comprising the same event. One configuration being a rectangle, this paper finds factorial moments of the number of these rectangles. Asymptotic behavior of the random variable defined is also investigated.

Keywords and phrases: rectangle, factorial moments, lattice.

1 Introduction

Natural disasters, ecological disorders and epidemiological patterns of a disease that randomly inflict human dwellings, forests and plant nurseries over space and time often engage a scientist's curiosity to understand how these events are statistically distributed. For events that occur with same probability the concept of join count statistics is generally found useful. The number of these counts becomes a random variable of interest in situations specially where the locations define a lattice. For instance, fruit trees are often grown in this fashion for the purpose of improving efficiency of production but the risk of disease is there. Experiments with artificially generated lattices are conducted to examine the effect of soil infection and the infection patterns that it creates in trees. A disease has binary nature, its presence or absence, as it is the case of a single binomial trial where an event may or may not occur. A join occurs when two adjacent trees are simultaneously infected by disease in a horizontal, vertical or some other manner.

Single binominal trials find their application in scientific inquiries where a trial results in some specified event with a constant probability. A researcher may be interested in the number of these events that happen randomly when the trial is independently repeated.

Suppose that we have a set of n^2 locations of a square lattice, arranged in n rows and n columns, and that binomial trials occur simultaneously at all these locations. If each trial results in a specified event E with some probability, various configurations are likely to emerge. A configuration of these events and so its probability, often becomes a subject of interest. Moran (1948), Fuchs and David (1965) and Memon and David (1968) for instance, study the distribution of various patterns that arise in a similar situation based on single binomial trials. We investigate here the probability distribution of the number of particular configuration which we call here a rectangle of events and define it below with reference to the above mentioned situation.

Let $\phi_{i,j}$ denote j^{th} link in i^{th} row with a value = 1 (if the event E occurs at locations j and $j+1$ in i^{th} row) and zero otherwise. Thus the condition $\phi_{i,j} = \phi_{i+1,j} = 1$ entails a rectangle with the event E that occurs at the locations (i,j) , $(i,j+1)$, $(i+1,j)$, $(i+1,j+1)$ and zero otherwise. Let X be the number of rectangles that arise when independent Bernoulli trials materialize simultaneously at n^2 locations of the lattice. The use of Memon and David (1968) Theorem in Appendix will be made to find the factorial moments of X in this paper.

2 Factorial Moments of the Random Variable X

In this section we find the first three factorial moments.

2.1 First Factorial Moment

A particular rectangle occurs when $\phi_{i,j} = \phi_{i+1,j} = 1$, for $i,j = 1,\ldots,n-1$; and zero otherwise. For $r = 1$, in Memon and David Theorem we have the following first factorial moment

$$\mu_{[1]} = \sum_k p_k,$$

where p_k is the probability of a particular rectangle. The probability p_k of a particular rectangle is p^4 and the total number of such possible rectangles are

$$\sum_{i,j,r,s=1}^{n-1} \phi_{ij}\phi_{rs}$$

which simplifies to

$$\sum_{i,j=1}^{n-1} \phi_{ij}\phi_{i+1j} = (n-1)^2$$

that is

$$\mu_{[1]} = (n-1)^2 p^4. \tag{2.1}$$

2.2 Second Factorial Moment

The second factorial moment follows from $r = 2$, in the Memon and David Theorem, that is,

$$\mu_{[2]} = 2! \sum prob(\text{two particular rectangles})$$

where $\sum$ is carried over all possible sets of two rectangles in the $n \times n$ lattice. Since two rectangles may have a common side, a common corner, or non-contiguous with probabilities p^6 , p^7 and p^8 respectively. It follows that the above moment is expressed as

$$2!(a_{2,6}p^6 + a_{2,7}p^7 + a_{2,8}p^8). \tag{2.2}$$

We find below the coefficients of probabilities p^6, p^7 and p^8.

The coefficient $a_{2,6}$: To calculate this coefficient we consider the following model for two adjacent rectangles in a row $\phi_{i,j} = \phi_{i,j+1} = \phi_{i+1,j} = \phi_{i+1,j+1} = 1$, for $i = 1, \ldots, n-1$; $j = 1, \ldots, n-2$; and zero otherwise. Under these conditions the number of these possibilities can be obtained from

$$\sum_{i_1,\ldots,i_8} \phi_{i_1,i_2}\phi_{i_3,i_4}\phi_{i_5,i_6}\phi_{i_7,i_8}$$

over $i_1, \ldots, i_8 = 1, \ldots, n-1$; which simplifies to

$$n_1 = (n-1)(n-2).$$

Similarly the number of possibilities for column-wise two adjacent rectangles is

$$n_2 = (n-1)(n-2).$$

Hence,

$$a_{2,6} = n_1 + n_2 = 2(n-1)(n-2).$$

The coefficient $a_{2,7}$: To calculate this coefficient we consider the following model for the two diagonally ($> 90°$) adjacent rectangles with a common corner, i.e.,

$$\phi_{i,j} = \phi_{i,j+1} = \phi_{i+1,j+1} = \phi_{i+2,j+1} = 1,$$

for $i, j = 1, \ldots, n-2$ and zero otherwise. The number of these configurations simplifies to

$$\sum_i \sum_j \phi_{i,j}\phi_{i,j+1}\phi_{i+1,j+1}\phi_{i+2,j+1}$$

that is,

$$n_3 = (n-2)^2.$$

Similarly the number of two diagonally adjacent ($< 90°$) rectangles with a common corner is

$$n_4 = (n-2)^2.$$

so

$$a_{2,7} = n_3 + n_4 = 2(n-2)^2.$$

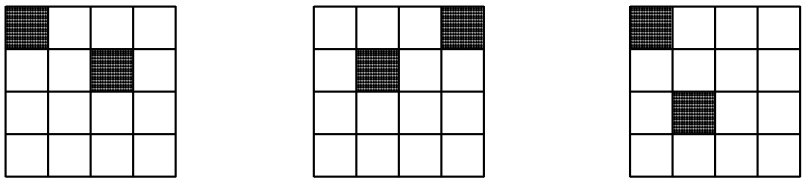
The coefficient $a_{2,8}$: For this coefficient we have the following model for the two non-contiguous rectangles in a row: $\phi_{i,j} = \phi_{i+1,j} = \phi_{i+1,m} = \phi_{i+1,m} = 1$ for $i = 1, \ldots, n-1$; $j = 1, \ldots, n-3$; $m = 3, \ldots, n-1$; $m - j \geq 2$; and zero otherwise. The number of these possibilities in the lattice can be shown as

$$n_5 = \frac{(n-1)(n-2)(n-3)}{2}.$$

Similarly the number of possible column-wise two non-contiguous rectangles in the lattice

$$n_6 = \frac{(n-1)(n-2)(n-3)}{2}.$$

The model, for horizontally appearing non-contiguous rectangles with one or more gaps in the adjacent rows (as indicated in Figure 1(a)), is $\phi_{i,j} = \phi_{i+1,j} = \phi_{i+1,k} = \phi_{i+1,k} = 1$ for $i = 1, \ldots, n-2; j = 1, \ldots, n-3; k = 3, \ldots, n-1$; $k - j \geq 2$; and zero otherwise.



(a) (b) (c)

Figure 1:

The number of possible arrangements for Figure 1(a) is given as:

$$n_7 = \frac{(n-2)^2(n-3)}{2}.$$

By symmetry, the number of arrangements for Figure 1(b) is:

$$n_8 = \frac{(n-2)^2(n-3)}{2}.$$

Figure 1(c) represents all situations involving non-adjacent rectangles with one or more horizontal and vertical gaps. The model for this situation is $\phi_{i,j} = \phi_{i+1,j} = \phi_{k,m} = \phi_{k+1,m} = 1$, for $i = 1, \ldots, n-2; j = 1, \ldots, n-1; k = 3, \ldots, n-1; m = 2, \ldots, n-1$; $k \geq i+2$; $m \geq j+1$, and zero otherwise.

We have the number of these situations

$$n_9 = \frac{(n-1)(n-2)^2(n-3)}{2}.$$

Hence

$$\begin{aligned} a_{2,8} &= n_5 + n_6 + n_7 + n_8 + n_9 \\ &= (n-1)(n-2)(n-3) + (n-2)^2(n-3) + \frac{(n-1)(n-2)^2(n-3)}{2}. \end{aligned}$$

2.3 Third Factorial Moment

The third factorial moment follows from $r = 3$ in the theorem referred above. Here,

$$\mu_{[3]} = 3! \sum prob(\text{three particular rectangles}),$$

where the summation $\sum$ is carried over all possible sets of three rectangles with probabilities p^8, p^9, p^{10}, p^{11} and p^{12} respectively, depending on how these rectangles emerge. The three rectangles may have one or more common corners, one or more common sides, or they may be non-adjacent. The third moment can be expressed as

$$3!(a_{3,8}p^8 + a_{3,9}p^9 + a_{3,10}p^{10} + a_{3,11}p^{11} + a_{3,12}p^{12}). \tag{2.3}$$

As done in the second factorial moment, we specify their locations and formulate a model involving the variables $\phi_{i,j}$ and enumerate the possibilities using the product $\prod_{i}^{6} \phi_{m_i,n_i}$, where m_i, n_i take values from $1, \ldots, n-1$.

The coefficient $a_{3,8}$: To obtain this coefficient we consider all three rectangles adjacently appearing in a row with probability p^8. Since in this situation the adjacent rectangles have a common vertical side we set up the following model; $\phi_{i,j} = \phi_{i,j+1} = \phi_{i,j+2} = \phi_{i+1,j} = \phi_{i+1,j+1} = \phi_{i+1,j+2} = 1$ for $i = 1, \ldots, n-1; j = 1, \ldots, n-3$; and zero otherwise. The number of these possibilities can be thus determined from the simplified expression

$$\sum_i \sum_j \phi_{i,j}\phi_{i+1,j}\phi_{i,j+1}\phi_{i+1,j+1}\phi_{i,j+2}\phi_{i+1,j+2}$$

and so

$$n_{10} = (n-1)(n-3). \tag{2.4}$$

By symmetry, the number of possible cases when three rectangles appear adjacently in columns has to be same as Equation 2.3, that is,

$$n_{11} = (n-1)(n-3). \tag{2.5}$$

Three rectangles one of which has a common vertical side with the second rectangle on the right and a common horizontal side with the lower third may be described by the model; $\phi_{i,j} = \phi_{i+1,j} = \phi_{i,j+1} = \phi_{i+1,j+1} = \phi_{i+2,j} = 1$, for $i = 1, \ldots, n-2; j = 1, \ldots, n-2$; and zero otherwise. The number of such events is found as

$$n_{12} = (n-2)^2. \tag{2.6}$$

The clock-wise rotation of three rectangles considered above produces three similar situations in addition. By symmetry the number of possibilities correspondingly is the same as 2.6, that is,

$$n_{12} = n_{13} = n_{14} = n_{15}. \tag{2.7}$$

Consequently, the coefficient of p^8 from Equations 2.4, 2.5, 2.6 and 2.7 can be simplified to

$$a_{3,8} = 6n^2 - 24n + 22.$$

The other coefficients in Equation 2.3 determined similarly by using the above approach are

$$\begin{aligned}
a_{3,9} &= 8n^2 - 40n + 48.\\
a_{3,10} &= 12n^3 - 96n^2 + 246n - 198.\\
a_{3,11} &= n^4 - 2n^3 - 39n^2 + 172n - 192.\\
a_{3,12} &= \frac{1}{6}(n^6 - 6n^5 + 6n^4 - 68n^3 - 725n^2 + 172n - 192).
\end{aligned}$$

Remark 1. With an $n \times n$ lattice the usual approach for finding moments of the random variable X requires the knowledge of its distribution. As n becomes larger the derivation of this distribution becomes more cumbersome. On the contrary, the use of theorem simplifies the calculation of moments whatever n may be. Even for n as small as 3 it can be seen that it is not simple to obtain the following:

$$\begin{aligned}
p(0) &= q^9 + 9q^8p + 36q^7p^2 + 84q^6p^3 + 122q^5p^4 + 106q^4p^5 + 48q^3p^6 + 10q^2p^7 + qp^8\\
p(1) &= 4q^5p^4 + 20q^4p^5 + 32q^3p^6 + 12q^2p^7\\
p(2) &= 4q^3p^6 + 14q^2p^7 + 4qp^8\\
p(3) &= 4qp^8\\
p(4) &= p^9,
\end{aligned}$$

where $q = 1 - p$.

The moments of X are $E(X) = 4p^4$, $E(X^2) = 4p^4 + 8p^6 + 4p^7$ and $E(X^3) = 4p^4 + 24p^6 + 12p^7 + 24p^8$. But the above results easily follow from the factorial moments given in Equations 2.1, 2.2, 2.3 for $n = 3$.

2.4 Asymptotic Distribution of X

Assuming that $n^2p^4 \to \lambda$ as $n \to \infty$, and p is small, let us find the asymptotic moments of the random variable X defined above. The moments given in Equations 2.1, 2.2, 2.3 can be expressed as

$$\mu_{[i]} = (n^2p^4)^i \phi_i\left(\frac{1}{n}\right), \tag{2.8}$$

where $\phi_i(\frac{1}{n}) = 1 + \frac{a_{i1}}{n} + \frac{a_{i2}}{n^2} + \cdots$

In fact the Expression 2.8 holds for all factorial moments $i = 1, 2, \ldots$ as the maximum power of the term n^2p^4 appearing in the ith factorial moment turns to be i. Since $\phi_i(\frac{1}{n}) \to 1$ when $n \to \infty$, the ith factorial moment of X has asymptotically the values λ^i. That is, X has asymptotically a Poisson distribution with the parameter λ.

To illustrate it, suppose that a fruit plant is grown at each of n^2 locations of an $n \times n$ lattice. Let p be the probability that a disease infects each plant independently during some

time period. The number of rectangles could be an important variable of concern to a plant pathologist. For large n its distribution is approximately Poisson.

A Appendix

In their paper [2], Memon and David provide a result that facilitates a relationship between factorial moments and probabilities of specified events. They consider n possibly dependent events each of whose materialization is determined by a single binomial trial. Then the rth factorial moment of the number of materializing events is

$$\mu_{[r]} = \sum_{\binom{n}{r}} p(W)$$

where $p(W)$ denotes the probability of materialization of all events in a set of size r and the summation extends over all $\binom{n}{r}$ sets of size r.

Acknowledgement

The authors are grateful to the refree for his useful suggestions.

References

[1] Fuchs, C. E. and David, H. T. (1965). Poisson limits of multivariate run distributions. *Annals of Mathematical Statistics*, **36**(1), 215–225.

[2] Memon, A. Z. and David, H. T. (1968). The Distribution of lattice join counts. *Bulletin of the Institute of Statistical Research and Training*, **2**(2), 75–83.

[3] Moran, P. A. P. (1948). Interpretation of statistical maps. *Journal of the Royal Statistical Society, Series B*, **10**(2), 243–251.