Modeling the Correlation in Binary Sequences using Partial Exchangeability

Hanxiang Peng

Department of Mathematical Sciences
Indiana University-Purdue University at Indianapolis

Joint work with Stephine L. Keeton (Food and Drug Administration), Latonya C. Garner (Mississippi Valley State University), Gibson J. Rayner (Charleston Southern University), Xin Dang (University of Mississippi),

April 16, 2010

Outline

1. Motivating Examples
2. The Distribution of Partially Exchangeable Random Variables
3. Application I: Partially Exchangeable Models
4. Application II: Exchangeable Models

Cow data (Quintana and Newton, 1998)

The data consists of series of measurements of presence(1)/absence(0) about pathogen infection on cows. A pathogen is any disease producing micro-organism. Pathogen infection causes mastitis, which is inflammation of the mammary gland or breast. The presence or absence of a pathogen can be determined from examining glandular organs for mastitis. Measurements contained in the data set were collected in each of the four quarters of each cow's udder at 11 time periods after lactation. Many of these cows were measured on several consecutive lactation periods, which took about 1 year each. One period of each cow was randomly selected, and restriction was made to those cases where at least five observations were available.

Bladder Cancer Study (Quintana and Müller, 2004)

The study was conducted by the Veterans Administration Cooperative Urological Research Group (VACURG). State I bladder tumors can usually be completely removed by transurethral resection, that is performed through a tube (urethra) which conveys urine from the bladder to the outside. Unfortunately, many patients have multiple recurrences. In order to determine if recurrences of Stage I bladder cancer can be prevented, a randomized clinical trial was conducted. All patients had superficial bladder tumors when they entered the trial. These tumors were removed transurethrally and patients were assigned to one of the three groups: placebo, thiotepa, and pyridoxine (Vitamin B6). At subsequent follow-up visits, any recurrent tumors were removed and the treatment was continued.

Partial Exchangeability

- Indicators of bladder cancer and treatment:

$$B = \begin{cases} 1, & \text{cancer} \\ 0, & \text{normal} \end{cases} \qquad X = \begin{cases} 1, & \text{treatment} \\ 0, & \text{control} \end{cases}$$

- There are 73 patients. For each patient, a binary sequence $B_0, \ldots, B_m$ was observed, with max sequence length $m = 11$.
- $B_0 \ldots B_m$ are correlated. Let $\mathbf{t} = \{t_{00}, t_{01}, t_{10}, t_{11}\}$ be transition counts. For 0011100101, $t_{00} = 2, t_{01} = 3, t_{10} = 2, t_{11} = 2$.

Definition

Binary r.v.'s $B_0, B_1, \ldots$ are partially exchangeable if the joint distribution of any finite sub-sequence is invariant under permutations that leave the initial state and the transition counts $\mathbf{t}$ unchanged.

The Distribution of P. Exch. R.V. (Peng, et al., 2010)

Theorem

Suppose $B_0, B_1, \ldots$ are partially exchangeable and recurrent. Then

$$\begin{aligned} &\mathbb{P}(B_0 = b_0, \ldots, B_m = b_m) \\ &= \textstyle\sum_{j=0}^{t_{01}} \sum_{k=0}^{t_{10}} (-1)^{j+k} \binom{t_{01}}{j} \binom{t_{10}}{k} \lambda_{t_{00}+j, t_{11}+k}^{(b_0)}, \end{aligned} \tag{1}$$

where $\lambda_{j,k}^{(b_0)}$ is the joint probability that the transitions $0 \to 0$ and $1 \to 1$ simultaneously occur j and k times with initial state b_0, i.e,

$$\begin{aligned} &\lambda_{j,k}^{(b_0)} = \mathbb{P}\left(B_0 = b_0, (0 \to 0)^j, (1 \to 1)^k\right), \\ &0 \le j \le t_{00} + t_{01}, \; 0 \le k \le t_{11} + t_{10}. \end{aligned}$$

Proof. By Diaconis and Freedman (1980a, c).

The Distribution of Partially Exchangeable R.V.'s

- Link each $\lambda_{j,k}$ to $\boldsymbol{\beta}^\top X$ via an inverse link $h_{b_0,j,k}$:

$$\lambda_{j,k}^{(b_0)} = h_{b_0,j,k}(\boldsymbol{\beta}^\top X), \tag{2}$$

  where $\boldsymbol{\beta}$ is parameter of interest and X is a covariate vector.

- Substituting (2) in (1),

$$f(\mathbf{t}, \boldsymbol{\beta}) = \sum_{i=0}^{t_{01}} \sum_{j=0}^{t_{10}} (-1)^{i+j} \binom{t_{01}}{i} \binom{t_{10}}{j} h_{b_0, t_{00}+i, t_{11}+j}(\boldsymbol{\beta}^\top X),$$

- For f to be a probability mass function, $\left\{h_{b_0,j,k}(\boldsymbol{\beta}^\top X)\right\}$ has to be rectangular completely monotone (RCM):

$$(-1)^{r_1+r_2} \Delta_1^{r_1} \Delta_2^{r_2} h_{b_0,j,k}(\boldsymbol{\theta}) \geq 0, \tag{3}$$

  with $h_{0,0,0}(\boldsymbol{\theta}) + h_{1,0,0}(\boldsymbol{\theta}) = 1$ for all $\boldsymbol{\theta}$.

RCM links: Laplace Transforms

1. Laplace transform are RCM links:

$$h(s,t;\boldsymbol{\theta}) = \int_0^\infty \int_0^\infty \exp(-sx-ty)dH(x,y;\boldsymbol{\theta}) \quad s,t \in [0,\infty),$$

   where H is a probability measure on $[0,\infty)^2$.

2. Biga link: Laplace transform of bivariate gamma

$$h(s,t;\boldsymbol{\theta}) = \frac{(1+s)^{\theta_3-\theta_1}(1+t)^{\theta_3-\theta_2}}{(1+s+t+\rho st)^{\theta_3}}, \quad \boldsymbol{\theta} \in [0,\infty)^3 \times [0,1].$$

3. Kbiga link: Laplace transform of Kibble's bivariate gamma

$$h(s,t;\boldsymbol{\theta}) = \frac{(\theta_1\theta_2)^\nu}{((\theta_1+s)(\theta_2+t)-\rho st)^\nu}, \quad \boldsymbol{\theta} \in (0,\infty)^2 \times (0,1) \times (0,\infty).$$

RCM links: Existing Links

If $g(s,t;\boldsymbol{\theta})$ and $h(s,t;\boldsymbol{\theta})$ are RCM, then

- (LP) $wg+(1-w)h$ and $g*h$ are RCM.
- (C1) $h(\psi_1(s),\psi_2(t);\boldsymbol{\theta})$ is RCM, where $\psi_1(s)$ and $\psi_2(t)$ are positive functions with CM derivatives.
- (C2) $\varphi(g(s,t),h(s,t))$ is RCM, where $\varphi(u,v)$ is a nonnegative polynomial.

RCM Links: UCM links

- Peng, et al. (2009) proposed an unified approach for analyzing exchangeable binary data with applications to clinical and developmental toxicity studies.
- $B_1, \ldots, B_m$ are *exchangeable* if

$$\mathbb{P}(B_1 = b_1, \ldots, B_m = b_m) = \mathbb{P}(B_{\pi_1} = b_1, \ldots, B_{\pi_m} = b_m), \qquad (4)$$

  for every permutation $\pi_1, \ldots, \pi_m$ of $1, \ldots, m$.
- Let $Y = B_1 + \ldots + B_m$ be the total number of "successes". Then

$$\mathbb{P}(Y = y) = \binom{m}{y} \sum_{k=0}^{m-y} (-1)^k \binom{m-y}{k} \lambda_{y+k}, \; y = 0, 1, \ldots, m,$$

  where $\lambda_0 = 1$, $\lambda_k = \mathbb{P}(B_1 = 1, \ldots, B_k = 1)$ are the *marginal probabilities*.

RCM Links: UCM links

$\boldsymbol{\lambda} = \{\lambda_i\}$ $(\lambda_0 = 1)$ are *univariate completely monotone (UCM)*:

$$(-1)^k \Delta^k \lambda_i \geq 0, \quad i = 0, 1, ..., m, \tag{5}$$

where Δ is the difference operator: $\Delta\lambda_i = \lambda_{i+1} - \lambda_i$.

Table: **The Complete Monotone Links.**

Name	Link$(\theta = (\theta_1, \theta_2))$	Parameters
Ind-Bin	θ^t	$\theta \in (0, 1)$
MM-Bin	$\theta/(\theta + t)$	$\theta \in (0, \infty)$
Beta-Bin	$B(\theta_1 + t, \theta_2)/B(\theta_1, \theta_2)$	$\boldsymbol{\theta} \in (0, \infty)^2$
Gamma-Bin	$(1 + \theta_2 t)^{-\theta_1}$	$\boldsymbol{\theta} \in (0, \infty)^2$
Poisson-Bin	$\exp(\theta(e^{-t} - 1))$	$\theta \in (0, \infty)$
Normal-Bin	$2\exp((\sigma t)^2/2)(1 - \Phi(\sigma t))$	$\sigma^2 \in (0, \infty)$

Results of Cow data

Table: **Estimation of the Cow Data.**

Models	Parameter Estimates(s.d)	-loglik	P_{00}(s.d)	P_{11}(s.d)
Observed			.896	.967
$\text{Bin}_s(\theta_1) * \text{Ga}_t(\theta_2)$	$0.896(0.01)\ 0.108(0.01)$	503.8	$.896(.010)$	$.928(.009)$
$\text{MM}_s(\theta_1) * \text{MM}_t(\theta_2)$	$6.480(1.03)\ 25.95(3.44)$	507.0	$.866(.018)$	$.963(.005)$
$\text{Bin}_s(\theta_1) * \text{MM}_t(\theta_2)$	$0.896(0.01)\ 25.95(3.38)$	507.3	.**896**(.010)	.**963**(.005)
$\text{Bin}_s(\theta_1) * \text{Ga}_t(\theta_2, \theta_3)$	$0.896(0.01)\ 0.191(0.06)\ 0.315(0.16)$	**501**.2	$.896(.010)$	$.949(.010)$
$\text{Kbg}_{s,t}(\theta_1, \theta_2, \nu, \rho = 0)$	$5.900(1.02)\ 23.83(2.50)\ 0.928(0.19)$	507.0	$.864(.032)$	$.962(.008)$
$\text{Kbg}_{s,t}(\theta_1, \theta_2, \rho, \nu = 1)$	$6.480(1.04)\ 25.95(3.48)\ 0.000(0.00)$	507.0	$.866(.019)$	$.963(.006)$
$\text{Bin}_s(\theta_1, \nu) * \text{MM}_t(\theta_2)$	$0.876(0.03)\ 0.901(0.12)\ 25.94(3.90)$	506.6	$.876(.031)$	$.963(.005)$

The Results of Bladder Cancer Data

Table: **Estimated & Observed Transition Probabilities for the Bladder Cancer Data.**

	Control		Treatment	
	P_{00} (s.d.)	P_{11} (s.d.)	P_{00} (s.d.)	P_{11} (s.d.)
Observed	0.829	0.374	0.942	0.333
$\mathrm{Ga}_s(\alpha_1, \beta_1) * \mathrm{Ga}_t(\alpha_2, \beta_2)$	0.788(.036)	0.311(.057)	0.903(.027)	0.264(.087)
$\mathrm{Ga}_s(\alpha_1, \beta_1) * \mathrm{MM}_t(\alpha_2, \beta_2)$	0.788(.036)	0.362(.064)	0.903(.027)	0.307(.098)
$\mathbf{Bin}_s(\alpha_1, \beta_1) * \mathbf{MM}_t(\alpha_2, \beta_2)$	**0.829(.005)**	**0.361(.062)**	**0.943(.004)**	**0.320(.098)**
$\mathrm{MM}_s(\alpha_1, \beta_1) * \mathrm{MM}_t(\alpha_2, \beta_2)$	0.748(.036)	0.362(.064)	0.930(.016)	0.308(.094)
$\mathrm{Biga}_{s,t}(\alpha_1, \beta_1, \alpha_2, \beta_2, \theta_3, \rho)$	0.787(.036)	0.312(.058)	0.903(.027)	0.264(.087)
$\mathrm{Ga}_s(\alpha_1, \beta_1, \theta) * \mathrm{MM}_t(\alpha_2, \beta_2)$	0.785(.041)	0.362(.064)	0.901(.030)	0.307(.098)
$\mathrm{Kbiga}_{s,t}(\alpha_1, \beta_1, \alpha_2, \beta_2, \rho, v = 1)$	0.749(.036)	0.365(.064)	0.930(.016)	0.307(.094)
$\mathrm{Kbiga}_{s,t}(\alpha_1, \beta_1, \alpha_2, \beta_2, v, \rho = 0)$	0.753(.195)	0.412(.145)	0.926(.109)	0.351(.162)

Fitting the E2 Data.

The E2 data (Brooks *et al.* 1997) records fetal control mortality in mouse litters. There are 211 litters in total with litter sizes varying from small (as to 3) to large (as to 19), having the mean litter size 12.9 and the standard deviation 2.68. The proportion of dead fetuses is 0.110. Among the 211 litters, there are 135 litters which have at least one fetal fetus.

Table: **Fitting The E2 Data**. Observed: $(p, q) = (0.110, 0.640)$.

Models (npr)	$\hat{p}(s.d.)$	$\hat{\phi}(s.d.)$	$\hat{q}$	$-$2logL	AIC	BIC
Bin (1)	0.113(.006)	0.000		765.6	767.6	767.9
Correlated Bin (2)	0.131(.010)	0.073(.012)		720.5	724.5	731.2
Beta-Bin (2)	0.112(.009)	0.101(.017)	0.612	689.8	693.8	700.5
Two Bin (3)	0.111	0.114		682.4	688.4	698.5
Three Bin (5)	0.121	0.101		679.7	689.7	706.5
Beta-Bin with Bin (4)	0.135	0.189		680.2	688.2	701.6
Kuk's Q-power (2)	0.119	0.209	0.648	687.1	691.1	697.8
Kuk's P-power (2)	0.109	0.080	0.595	698.8	702.8	709.5
Gamma-Bin with $\theta_2 = 1$(1)	0.118(.0015)	0.191(.0009)	0.543	697.8	699.8	700.1
Gamma-Bin (2)	0.110(.0087)	0.093(.0181)	0.619	679.9	683.9	**684.5**
Inc. Gamma-Bin (4)	0.109(.0118)	0.101(.0301)	**0.633**	**675.2**	**683.2**	696.6
Piecewise-Flogit (1)	0.111(.0154)	0.112(.0362)	0.601	680.8	**682.8**	**683.1**
Piecewise-Flogit Power (2)	0.111(.0175)	0.112(.0411)	0.601	680.8	684.8	691.5
Inc. Beta-Bin (4)	0.110(.0055)	0.102(.0384)	0.632	676.4	684.4	697.8
Inc. A-Bin (2)	0.115(.0085)	0.096(.0172)	0.624	681.6	685.6	692.3

The estimated response probability $\hat{p}$, intra-litter correction $\hat{\phi}$, probability $\hat{q}$ of the affected litters, along with negative twice log-likelihood, AIC and BIC. The upper, middle, and lower table are from Brooks *e*t al. (1997) and Brooks (2001), Kuk (2005), and the proposed framework, respectively. The standard deviations are included when they are available. Highlighted are the optimal models.

Regression Analysis on the CD1 Data.

A developmental toxicology study conducted at the National Center for Toxicological Research. The study involves replicate experiments with 9 strains of female mice exposed to the herbicide 2,4,5-Trichlorophenoxyacetic acid. We use the data for the CD1 mice, which was analyzed by many authors, e.g., George and Bowman (1995) and Kuk (2004).

Regression Analysis on the CD1 Data.

A developmental toxicology study conducted at the National Center for Toxicological Research. The study involves replicate experiments with 9 strains of female mice exposed to the herbicide 2,4,5-Trichlorophenoxyacetic acid. We use the data for the CD1 mice, which was analyzed by many authors, e.g., George and Bowman (1995) and Kuk (2004).

Table: **Summary of The CD1 Data.**

Dose	nLitters	nImplants	avgLittersize	stdLittersize	nMalfs	pMalf
0	73	777	10.64	2.69	59	0.076
30	87	952	10.94	2.88	124	0.130
45	98	1124	11.47	2.93	338	0.301
60	76	806	10.61	2.99	390	0.484
75	44	482	10.95	2.23	372	0.772

Table: **The Estimates Under Various Models For The CD1 Data**. Link1: $(1+t)^{-\theta_1}$, Link2: $(1+\ln(1+t))^{-\theta_1}$, Link3: $(1+\theta_2 t)^{-\theta_1}$(Gamma-Bin), Link4: $(1+\theta_2\ln(1+t))^{-\theta_1}$(Gamma-log-Bin)

Model (npr)	α(s.d.)	β(s.d.)	-2logL	χ^2	AIC	BIC
Binomial (2)	-3.235(.113)	5.430(.217)	2295.0	1514.2	2298.9	2306.8
Beta-Bin (4)	0.433(.081)	3.788(.194)				
	0.503(.064)	-4.786(.118)	1464.8	336.43	1472.8	1488.5
GEE(Logit, Ex) (3)	-3.323(.205)	5.580(.438)		411.43		
Williams' (3)	-3.237(.231)	5.587(.442)		370.87		
Kuk's Q-power (3)	0.884(.071)	-0.525(.115)	1459.6	298.95	1465.6	1471.3
Link1 (2)	3.838(.174)	-4.712(.265)	1460.8	372.22	1464.8	1472.7
Link2 (2)	4.633(.259)	-5.447(.411)	1472.2	247.64	1476.2	1484.1
Link3 (3)	3.340(.161)	-4.082(.243)	1458.9	343.81	1464.9	1476.7
Link4 (3)	15.03(.759)	-18.16(1.16)	1450.1	312.63	1456.1	1467.9
aLink1+$(1-a)$Link2 (3)	4.116(.203)	-4.975(.315)	1445.8	310.53	1451.8	1463.6
aLink3+$(1-a)$Link4 (4)	7.884(.640)	-9.571(.878)	**1438.6**	338.84	**1446.6**	**1462.3**
Link1*Link2 (2)	2.119(.105)	-2.568(.161)	1459.4	309.66	1463.4	1471.3
Link3*Link4 (3)	2.404(.062)	-2.927(.292)	1458.4	326.75	1464.5	1476.2
Piecewise-Flogit (2)	6.792(.265)	-8.441(.393)	1501.8	510.81	1505.8	1513.7

References

1. H. Peng, X. Dang, X. Wang. The Partially Exchangeable Distribution. *Statist. Probabil. Lett.* In press, (2010).
2. F. Tan, G. J. Rayner, X. Wang, H. Peng. A Full Likelihood Procedure of Exchangeable Negative Binomials for Modelling Correlated and Overdispersed Count Data. *J. Statist. Plann. Inferr.* In press, (2010).
3. X. Dang, S. L. Keeton, H. Peng. A Unified Approach for Analyzing Exchangeable Binary Data with Applications to Clinical and Developmental Toxicity Studies. *Statist. Med.* 28: 2580-2604 (2009).
4. H. Peng, F. Tan, S. L. Keeton, X. Dang, Y. Wang Efficient Semiparametric Regresson in Exchangeable Models with Applications in Developmental Toxicity Studies. *Work in Progress*. (2010).

THANKS