Spatially correlated binary data modelling using generalized estimating equations with alternative hypersphere decomposition

Junjie Li and Jianxin Pan
Mathematical College, Sichuan University, China

1Email: dtljiljj@163.com

Abstract. For correlated binary data, generalized estimating Equation (GEE) approach is commonly used for inference of the mean parameters. When the associated correlation is of interest, e.g., in spatial data analysis, some traditional correlation structures such as compound symmetry and AR(1) were employed in the literature, which is not appropriate for the inherent correlation for correlated binary data. In this paper, we proposed another set of GEE implemented with alternative hypersphere decomposition (AHPC) to model the correlation parameters. An iterative algorithm is provided to estimate the mean and correlation parameters simultaneously. By utilizing the AHPC, we developed a parametrization for the correlation matrix that automatically ensures the order independence and non-negative definiteness. We also addressed the issue of asymptotic properties. For illustration, the proposed methods were used to fit a spatial bovine tuberculosis infection data in Ireland, aiming to analyse the influential factors which affect the infection of cattle and/or badgers.

1. Introduction
Correlated binary data are very common in practice especially in clinical trials and environmental research. Clinical trials often involve longitudinal studies, in which within-subject observations are taken repeatedly over time. Since repeated measurements concentrated from the same subject are not independent, it is necessary to analyse the within-subject correlation and in certain circumstances (e.g. in infectious disease studies) the correlation is of primary interest. Diggle [1] reviewed available methods to model longitudinal data. A parallel field is spatial data analysis, where the measurement time in longitudinal data analysis is replaced with location. In both longitudinal and spatial data analyses, correlated binary data are not uncommon and have to be modelled properly.

Regression models of the mean structure for longitudinal and spatial data were studied considerably in the literature. Liang and Zeger [2] proposed generalized estimating equations (GEE) to analyse such data even if they are not normally distributed. The method specifies a working structure to the within-subject covariance matrix in order to improve the estimating efficiency of the mean parameters in the longitudinal framework. Prentice [3] applied the GEE approach to modelling of correlated binary data, where one estimating equation was used to estimate the mean parameters and another estimating equation was employed to estimate the parameters in the working covariance structure. However, these approaches assume specific working covariance or correlation structures which may not always appropriate in practice. To resolve this constraint, joint model for the mean and covariance was proposed in [4] and [5] for normal responses, where Modified Cholesky decomposition (MCD) was used for modelling covariance matrices. Ye and Pan [6] modelled the...
mean and covariance structures, simultaneously, in the framework of GEE implemented with MCD. Pourahmadi [7] considered an alternative Cholesky decomposition (ACD) method and Zhang et al [8] proposed a mean-variance correlation model by applying hypersphere decomposition (HPC) to correlation matrices. However, all the aforementioned decompositions depend on chronologically order (e.g., time) of observations and are not appropriate for unordered observations such as spatial data. Recently, Li [9] proposed alternative hypersphere decomposition (AHPC) based method to model unordered observations but focused on application to continuous data such as normally distributed data.

In this paper we proposed two generalized estimating equations implemented with the AHPC to model the mean and correlation parameters specifically for correlated binary data, without any covariance structure specification. Due to the order-independent property of the AHPC, the proposed method can be used to handle spatially correlated binary data. In Section 2, we review the AHPC method first and then propose a joint GEE method for modelling the mean and correlation coefficients. We also used quasi-fisher scoring algorithm to solve the GEE and presented the algorithm in Section 3. Finally, the proposed method was applied to spatial bovine tuberculosis (TB) infection data analysis, aiming to assess the effects of various factors on TB infection for both cattle and badgers, and on the correlation of cattle herds and badger setts respectively.

2. Statistical models

2.1. Alternative hypersphere decomposition

Before introducing the AHPC method, we need to briefly review the HPC method [9]. The correlation matrix \( R = (\delta_{jk}) \) can be decomposed into \( R = T T' \) where \( T \) is a lower triangular matrix as bellow:

\[
T_i = \begin{pmatrix}
1 & 0 & 0 & \cdots & 0 \\
c_{i21} & s_{i21} & 0 & \cdots & 0 \\
c_{i31} & c_{i32}s_{i31} & s_{i32} & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
c_{im1} & c_{im1}s_{im1} & c_{im2}s_{im2} & \cdots & \prod_{l=1}^{m-1} s_{iml}
\end{pmatrix}
\]

where \( c_{jk} = \cos(\phi_{jk}) \) and \( s_{ijk} = \sin(\phi_{ijk}) \) are functions of certain angles \( \phi_{ijk} \), the \((j,k)\) th element of the matrix \( \Phi \), which has the functional relationship with the correlation coefficient as follows

\[
\delta_{jk} = \cos(\phi_{jk}) \prod_{l=1}^{k-1} \sin(\phi_{jl}) \sin(\phi_{dl}) + \sum_{l=1}^{k-1} \cos(\phi_{jl}) \cos(\phi_{dl}) \prod_{r=1}^{l-1} \sin(\phi_{jr}) \sin(\phi_{dr})
\]

This expression is complicated and also shows the origin of its order-dependence. More details including numerical examples can be found in [10], where the AHPC was proposed. The main idea of the AHPC is given as follows.

Within the AHPC, the correlation matrix again can be decomposed into \( R = T T' \), where \( t_j \) is the \( j \) th row vector of the matrix \( T \). We denote by \( \phi_{jk} \) as the angle between the two row vectors \( t_j \) and \( t_k \). In other words, \( \phi_{jk} = \angle t_j, t_k \), \( 1 \leq k < j \leq m \). A trigonometric functional relationship between \( \phi_{jk} \) and correlation coefficient can be connected through \( \delta_{jk} = \cos(\phi_{jk}) \). When changing the order of variables, the position of \( \phi_{jk} \) in the angle matrix \( \Phi \) changes as well. However, the resulting correlation coefficient remains the same [9].
2.2. Generalized linear models

The focus of this paper is on the modelling for spatially correlated binary data. Let $y_{ij}$ denote the $j$th measurement of the $i$th subject and $y_i = (y_{i1}, y_{i2}, \ldots, y_{im})$ be the responses vector of the $i$th subject. The expected value $E(y_i) = u_i = (u_{i1}, u_{i2}, \ldots, u_{im})$ is a $(m_i \times 1)$ vector representing the means. And $\text{var}(y_i) = \Sigma_i$ is a $(m_i \times m_i)$ covariance matrix of $y_i$. The covariance matrix can be written as $\Sigma_i = D_i^{1/2} R_i D_i^{1/2}$, where $D_i$ is a diagonal matrix with the same diagonal elements as $\Sigma_i$, that is, $\text{Var}(y_{ij}) = u_{ij}(1-u_{ij})$, and $R_i$ is the correlation matrix with the elements $\text{Corr}(y_{ij}, y_{ik}) = \delta_{jk}$. In general, the matrices $R_i$ are supposed to be positive definite.

For the correlated binary data, we aim to model the mean and the correlation coefficient according to explanatory variables. By using the AHPC, this reduces to model the mean and the angle, simultaneously. Since the data are binary, we propose the models below for the mean and the angle,

$$\log(\frac{u_{ij}}{1-u_{ij}}) = x_{ij}' \beta \quad \phi_{ijk} = \arctan(w_{ijk}' \gamma) + \frac{\pi}{2}$$

(1)

where $x_{ij}$, $w_{ijk}$ are the $(p \times 1)$ and $(q \times 1)$ vectors of covariates, and $\beta$, $\gamma$ are the relevant regression parameters. The covariates $x_{ij}$ and $w_{ijk}$ may include basic covariates, GIS coordinates, and polynomials of time as well. For instance, in longitudinal data analysis the responses $y_i$ are measured at multiple time points which can be denoted by a $m_i \times 1$ vector as $\tilde{t}_i = (\tilde{t}_{i1}, \tilde{t}_{i2}, \ldots, \tilde{t}_{im})$. When using polynomial of time modelling the mean and the angle, the covariates may have the following formula

$$x_{ij} = (1, \tilde{t}_{i1}, \tilde{t}_{i2}^2, \ldots, \tilde{t}_{iP-1})'$$

$$w_{ijk} = (1, |\tilde{t}_{ij} - \tilde{t}_{ik}|, |\tilde{t}_{ij} - \tilde{t}_{ik}|^2, |\tilde{t}_{ij} - \tilde{t}_{ik}|^q)'$$

(2)

For the multicentral spatial data, $x_{ij}$ is the basic covariates, and $w_{ijk} = \|\text{Location}_{ij} - \text{Location}_{ik}\|$ may be taken as the absolute location distance or the GIS coordinate between the location $j$ and location $k$ for the center $i$, is case of within-centre correlation only relies on the locations.

2.3. Generalized estimating equations

Aiming at modelling the mean and correlation structure and estimating the parameters in Equation (1), we design two generalized estimating equations as follow, separately:

$$S_1(\beta) = \sum_{i=1}^{n} \left( \frac{\partial u_{ij}'}{\partial \beta} \Sigma_i^{-1}(y_i - u_i) \right)$$

$$S_2(\gamma) = \sum_{i=1}^{n} \left( \frac{\partial \text{svec}'(R_i)}{\partial \gamma} W_i^{-1}(\text{svec}(r_ir_i') - \text{svec}(R_i)) \right)$$

(3)

where $\frac{\partial u_{ij}'}{\partial \beta}$ is the $(p \times m_i)$ matrix with $j$th column $\frac{\partial u_{ij}'}{\partial \beta} = u_{ij}(1-u_{ij})x_{ij}$. And $r_i$ in $S_2(\gamma)$ are the $(m_i \times 1)$ vectors with $j$th components being the standardised residual

$$r_{ij} = \frac{y_{ij} - u_{ij}}{\sqrt{u_{ij}(1-u_{ij})}}$$

In (3) $\text{svec}(r_ir_i')$ represents the $m_i(m_i-1)/2$ -dimensional vector which vectorises the lower triangular elements of the matrix $r_ir_i'$, but excluding its diagonal elements, through column by column.
This applies to the notation \( \text{svec}(R_i) \) too. Note that by the definition we must have \( E(\text{svec}(rr')) = \text{svec}(R_i) \). Thus \( \delta \text{svec}'(R_i) / \partial \gamma \) is the \( q \times (m_i (m_i - 1) / 2) \) matrix with the element

\[
\frac{\partial \delta_{ik}}{\partial \gamma} = -\sin(\phi_{ik}) \cdot w_{ijk}^2
\]

In Equation (3), the covariance matrix of \( \text{svec}(rr') \) is \( W_i \), which is difficult to calculate. In spirit of GEE [2], we can use a sandwich 'working' structure \( W_i = A_i^{1/2} G_i(\rho) A_i^{1/2} \) to approximate the covariance matrix \( W_i \), where \( A_i \) is the \( m_i (m_i - 1) / 2 \)-dimensional diagonal matrix consisting of the variance of the vector \( \text{svec}(rr') \). Prentice [3] provided the variance of \( r'_i r'_k \) as

\[
\text{var}(r'_i r'_k) = 1 + (1 - 2u_{ij})(1 - 2u_{ik})(u_{ij}u_{ik} (1 - u_{ij})(1 - u_{ik}))^{-1} \delta_{jk} - \delta_{ik}^2
\]

The correlation matrix of the \( \text{svec}(rr') \) is approximated by \( G_i(\rho) \) which may depend on a new parameter \( \rho \). Commonly used structures for \( G_i(\rho) \) include Compound symmetry (CS), order-1 moving average (MA(1)) and order-1 autoregressive model (AR(1)). The inverses of such commonly used matrices have explicit forms so that their calculations are affordable even if the dimension of \( G_i(\rho) \) is very high. Note this method guarantees consistent estimates of parameters in 1st moment and 2nd moment without assuming any covariance structure. Pan [11] introduced the quasi-likelihood under the independence model criterion (QIC) for model selection within framework of GEE, which is defined by

\[
\text{QIC}(R) = -2\Psi(\hat{\beta}(R); I) + 2\text{trace}(\hat{Q}_i \hat{V}_r)
\]

where the quasi-likelihood

\[
\Psi(\hat{\beta}(R); I) = \sum_{i=1}^n \sum_{j=1}^{m_i} Q\left( \hat{\beta}(R), \hat{\phi}; \{y_{ij} \} \right) \text{ with } Q\left( u, \hat{\phi}; y \right) = \int_{-\infty}^{\infty} \left( (y - t) / \hat{\phi} V(t) \right) dt,
\]

in which \( \hat{\phi} \) is the estimate of dispersion parameter. It’s easy to verify that the \( S_i(\beta) \) in Equation (1) equivalent to \( \partial \Psi(\hat{\beta}(R); I) / \partial \beta \), \( \Omega_i \) can use the empirical estimator \( \hat{\Omega}_i = -\hat{\delta}^2 \Psi(\beta; I) / \partial \beta \beta' |_{\beta = \beta} \) to estimate. \( \hat{V} \) is the approximation of the covariance of \( \hat{\beta} \). The detailed expressions may refer to [2].

3. Estimators of parameter

The estimator of \( \beta \), \( \gamma \) satisfy the equations:

\[
S_1(\beta) = 0 \quad S_2(\gamma) = 0
\]

where \( S_1(\beta) \) and \( S_2(\gamma) \) are given in Equation (3). We use the Fisher-Score algorithm to get the estimator. We first calculate the quasi-Fisher information matrix \( I_\beta \) and \( I_\gamma \), which are the minus expectations of derivative of the score function \( S(\theta)' = (S_1'(\beta), S_2'(\gamma)) \) with respect to \( \theta \), where \( \theta = (\beta', \gamma')' \). We then can prove that the \( I_\theta \) is block-diagonal, that is \( I_\theta = \text{diag}(I_\beta, I_\gamma) \), where

\[
I_\beta = \sum_{i=1}^n \left( \frac{\partial u_i'}{\partial \beta} \Sigma_i^{-1} \left( \frac{\partial u_i'}{\partial \beta} \right)' \right) = \sum_{i=1}^n X_i' D_i^{-1/2} R_i^{-1} D_i^{1/2} X_i
\]

\[
I_\gamma = \sum_{i=1}^n \left( \frac{\partial \text{svec}'(R_i)}{\partial \gamma} \right) W_i^{-1} \left( \frac{\partial \text{svec}'(R_i)}{\partial \gamma} \right)'
\]

When \( \Sigma_i \) is given, the estimator of mean parameter \( \beta \) can be updated by
where \( \tilde{y}_i = (y_i - u_i) + (\partial u'_i / \partial \beta) \beta \). Then given \( \beta \), we can update correlation coefficient \( \gamma \) by

\[
\hat{\gamma} = \left\{ \sum_{i=1}^{n} \frac{\partial \text{svec}'(R)}{\partial \gamma} W_i^{-1} \text{svec}'(R_i) \right\}^{-1} \left\{ \sum_{i=1}^{n} \frac{\partial \text{svec}'(R)}{\partial \gamma} W_i^{-1} \left[ \text{svec}(r_i - R_i) + \frac{\partial \text{svec}(R_i)}{\partial \gamma} \right] \right\}
\]  

The Equations (6) and (7) suggest that that the estimators of parameters can be viewed as the weighted generalized least square estimators.

4. Analysis of the bovine tuberculosis infection data

In Ireland and the UK, cattle are infected by bovine tuberculosis (TB), a kind of disease that causes big problems for farmers. There is evidence that wild life badgers should be responsible for the spread of TB in cattle and perhaps vice versa. The data analysed below were derived from the Four Area Project, a formal project devoting to badger removal research performed in four counties in Ireland in 1997-2002, but we focused on Cork area for illustration, which has approximately 400 km². The scientists used the GIS coordinate to record the spatial location of the sett where the badger was captured. As for cattle, the GIS of the centroid of the main land where the cattle herd located was used. A herd is recorded as TB infected if any cattle of the herd is tested positive. As for the badgers, a sett is infected if any badger in it tests positive. Further details could be found in [11]. We intend to estimate the parameters of means and correlation coefficients which describe the badger setts and cattle herds respectively.

There were a total of 251 setts of badgers, 114 (41.5%) were TB positive while there were 417 cattle herds with 51 (12.2%) TB positive. For badgers there was one covariate: sett size. For cattle herds there were two covariates: previous history of infection in the herd (0 stands for negative and 1 positive) and log (herd size). The model of mean for badger data has the following expression
and for the cattle data, the model of mean takes the form of

$$g(E(Y_i)) = \beta_{c0} + \beta_{c1} \times ph + \beta_{c2} \times \logsize$$

where $g(.)$ is the logistic link function, $\beta$s are unknown parameters to be estimated, size denotes sett size, $ph$ denotes previous history for infection and $\logsize$ is the log of herd size. As for the correlation coefficients the models can be expressed as $\phi_{ij} = \arctan(w_{ij}\gamma) + \pi / 2$, where the covariates $w_{ij}$ is the distance between location$_i$ and location$_j$ of the badger setts or the cattle herds. Here we use the Euclidean distance

$$w_{ij} = \left[ (x_j - x_i)^2 + (y_j - y_i)^2 \right]^{1/2},$$

where $x_i$ and $y_i$ are the corresponding GIS coordinates of herd or sett. We then used the proposed joint GEEs to fit the badger data and cattle data. When estimating the correlation parameters, we took the compound symmetry (CS) and AR(1) structures by varying the parameter $\rho$ involved in $W_{ij} = A^{1/2} G_i(\rho) A^{1/2}$, the working covariance structure of $svec(r_{ij})$. We found that the resulting parameter estimators are very similar whenever using CS or AR(1) for $G_i(\rho)$. We only present the results for the case of AR(1) to improve the space efficiency.

**Table 1.** Badger Data Analysis. Parameter estimators and the standard errors as $\rho$ in $G_i(\rho)$ varies.

| $\rho$ | GLM | $\rho = 0$ | $\rho = 0.1$ | $\rho = 0.3$ | $\rho = 0.5$ | $\rho = 0.7$ | $\rho = 0.9$ |
|-------|-----|-----------|-----------|-----------|-----------|-----------|-----------|
| $\beta_{b0}$ | -1.2231 | -1.1818 | -1.1804 | -1.175 | -1.1614 | -1.1214 | -1.1919 |
| ste($\beta_{b0}$) | 0.2559 | 0.2211 | 0.2196 | 0.2135 | 0.1963 | 0.1182 | 0.2252 |
| $\beta_{b1}$ | 0.4636 | 0.4638 | 0.4638 | 0.464 | 0.4645 | 0.4665 | 0.4666 |
| ste($\beta_{b1}$) | 0.1015 | 0.1024 | 0.1024 | 0.1025 | 0.1028 | 0.1035 | 0.1024 |
| $\gamma$ | 0.00337 | 0.0035 | 0.00399 | 0.00528 | 0.00949 | 0.015 | 0.015 |
| ste($\gamma$) | 0.00357 | 0.00393 | 0.00475 | 0.00575 | 0.00669 | 0.00548 |
| QIC | 364.706 | 364.649 | 364.438 | 363.944 | 362.834 | 365.214 |

**Table 2.** Cattle Data Analysis. Parameter estimators and the standard errors as $\rho$ in $G_i(\rho)$ varies.

| $\rho$ | GLM | $\rho = 0$ | $\rho = 0.1$ | $\rho = 0.3$ | $\rho = 0.5$ | $\rho = 0.7$ | $\rho = 0.9$ |
|-------|-----|-----------|-----------|-----------|-----------|-----------|-----------|
| $\beta_{c0}$ | -2.549 | -2.7234 | -2.7233 | -2.7226 | -2.7192 | -2.7048 | -2.6627 |
| ste($\beta_{c0}$) | 0.2262 | 0.1216 | 0.1217 | 0.1222 | 0.1234 | 0.1268 | 0.1293 |
| $\beta_{c1}$ | 0.8396 | 0.8614 | 0.8605 | 0.8571 | 0.8511 | 0.8369 | 0.8029 |
| ste($\beta_{c1}$) | 0.3238 | 0.3341 | 0.3341 | 0.334 | 0.3336 | 0.3311 | 0.3237 |
| $\beta_{c2}$ | 0.9378 | 1.021 | 1.0218 | 1.0238 | 1.024 | 1.0123 | 0.9598 |
| ste($\beta_{c2}$) | 0.225 | 0.2366 | 0.2369 | 0.2379 | 0.2403 | 0.2472 | 0.2538 |
| $\gamma$ | 0.00388 | 0.00418 | 0.00532 | 0.00796 | 0.01487 | 0.02467 |
| ste($\gamma$) | 0.00204 | 0.00224 | 0.00269 | 0.00319 | 0.00356 | 0.00277 |
| QIC | 567.244 | 567.209 | 567.137 | 567.221 | 568.1918 | 572.409 |

Table 1 shows the output of QIC, mean and correlation coefficient parameter estimators (and their standard errors) for badger sett data using GLM and joint GEE using six different values of $\rho$ in AR(1). The standard errors of $\beta$s and $\gamma$s could be calculated from the quasi-fisher information.
matrix $I_\beta$ and $I_\gamma$. From Table 1 we can see that different choices of $\rho$ have little effect on the estimator of $\beta$, implying that the GEE estimators of parameters are robust against distinct structure of $G_\rho(\rho)$. In terms of the value of QIC, the optimal choice of $\rho$ seems to be $\rho = 0.7$. And the estimated correlation coefficient ranges from 0 to 0.007. Table 2 gives the estimating results of the QIC, mean and correlation coefficient parameters (and their standard errors) for cattle data. In terms of QIC, the optimal value is $\rho = 0.3$. In this case, the estimated parameters for ph and logsize are 0.8571 and 1.0238, respectively, indicating that previous infection and herd size both increase the risk of bovine tuberculosis infection. The estimated correlation coefficient ranges from 0 to 0.003. Although $\gamma$ is very small, the value of the estimator is larger than the twice standard error, meaning the parameter $\gamma$ is statistically significant.

The ste($\beta_0\rho$) by GLM is larger than that by the joint GEE. This is because GEE method reduces the standard error of the estimator, which is the merit of GEE method. We specify a covariance structure to the 4th moment, rather than the 2nd moment. The latter is the ordinary GEE method. In summary, our method guarantees the consistent estimate of parameters in 1st moment and 2nd moment without any covariance structure specification to the 2nd moment.

5. Conclusions
Correlated binary data are very common in many practices especially in clinical trial and environmental sciences. It is difficult to model such data due to insufficient data information. In the meantime, of interest is to model not only the mean but also the correlation for correlated binary data in terms of explanatory variables.

By using the alternative hypersphere decomposition, we transformed the correlation matrix to the angle matrix. We then proposed two generalized estimating equations to estimate the mean and correlation coefficients, simultaneously, for correlated binary data. The proposed method is able to handle not only longitudinal binary data but also spatial binary data. The key point is that the alternative hypersphere decomposition is order-independent. The resulting estimator of correlation matrix for correlated binary data could be ensured the nonnegative definiteness.

References
[1] Diggle P, Diggle P J, Heagerty P, Heagerty P J, Liang K Y and Zeger S 2002 Analysis of longitudinal data Oxford University Press
[2] Liang K Y and Zeger S L 1986 Longitudinal data analysis using generalized linear models. *Biometrika* 73(1) 13-22
[3] Prentice R L 1988 Correlated binary regression with covariates specific to each binary observation *Biometrics* 1033-1048
[4] Pourahmadi M 1999 Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation *Biometrika* 86(3) 677-690
[5] Pourahmadi M 2000 Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika* 87(2) 425-435
[6] Ye H and Pan J 2006 Modelling of covariance structures in generalised estimating equations for longitudinal data *Biometrika* 93(4) 927-941
[7] Pourahmadi M 2007 Cholesky decompositions and estimation of a covariance matrix: orthogonality of variance–correlation parameters *Biometrika* 94(4) 1006-1013
[8] Zhang W, Leng C and Tang C Y 2015 A joint modelling approach for longitudinal studies *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77(1) 219-238
[9] Li Q 2018 Covariance modelling. PhD thesis, University of Manchester, UK
[10] Pan W 2001 Akaike’s information criterion in generalized estimating equations *Biometrics* 57(1) 120-125
[11] Griffin J M, Williams D H, Kelly G E, Clegg T A, O’boyle I, Collins J D and More S J 2005 The impact of badger removal on the control of tuberculosis in cattle herds in Ireland *Preventive veterinary medicine* 67(4) 237-266