Independence Working Correlation Only: GEE for Quadratic Exponential Logistic Regressions

Abstract

For modeling correlated binary variables, the quadratic exponential distribution (Cox & Wermuth (1994)) is an educated choice. This distribution has not been popularized because it potentially requires massive computation when solving maximum likelihood estimates. However, the corresponding fully conditional distribution has the form of logistic regression, which motivates us to borrow the generalized estimating equation Liang & Zeger (1986) approach for estimation. We prove that the independence working covariance is the only valid one among popularly applied working covariances, although dependency between the fully conditional variables is expected. This counter-intuitive conclusion is because we aggregate all the fully conditional distributions but not the marginal distributions, which are generally required for the generalized estimating equation approach. With the proposed quadratic exponential logistic regression model, we explored the relationships among course aims based on the survey of an English elite course and investigated the influential factors for writing non-consensual opinions among grand justices.

Keywords: asymmetric Ising model; fully conditional distribution; network data

1 Introduction

This paper explores the factors that influence the strength of correlations among binary responses. In the context of binary response regression problems, three educated models are popular: generalized linear mixed models (GLMM; Breslow & Clayton (1993)), generalized estimating equations (GEE; Liang & Zeger (1986)), and quadratic exponential logistic regression (QELR) originated from Cox (1972), Cox & Wermuth (1994) and the references therein. Among these, GLMM and GEE have become the standard solutions. However, they differ in their foundational aspects, subject-specific and population-average, respectively, which lead to incomparable parameters. When the focus shifts to incorporating covariates into the correlation structure, a useful solution is found in the second-order GEE (Crespi et al. 2009). The QELR also has the potential to model correlations but receives less attention due to its computation complexity. It is worth noting that the QELR's fully conditional distribution takes the form of logistic regression, allowing it to be solved using GEE. This paper considers the QELR for correlation modeling and uses GEE as a solution.

Modeling graphs or social networks is another important application with multiple correlated binary responses. Holland & Leinhardt (1981) introduced the log-linear model, and Anderson et al. (1999) is a primer on modeling networks with log-linear related models. The log-linear model with main effects and two-way interaction effects is exactly a quadratic exponential binary distribution (QEBD; Cox & Wermuth (1994)). Moreover, originating from physics and serving as useful for social networks, the Ising model, e.g., Preston (1974), is also a special case of the QEBD. Network data usually consists of many correlated binary variables with only one replicate. However, in this paper, we consider the cases with a small number of correlated binary variables with many replicates. Thus, several

successful strategies that work for large network data, such as Ravikumar et al. (2010) and Anandkumar et al. (2012), are no longer applicable here. We refer to Strauss & Ikeda (1990), Anderson et al. (1999), and Zhao & Prentice (1990) for a review of fundamentals in solving Ising-like models.

This paper presents two contributions. Firstly, we represent the QEBD in a quadratic form. This transformation enables us to express subsequent formulas using matrix operations, significantly aiding the computing efficiency. Secondly, we highlight a GEE example that necessitates the independence working covariance. A similar statement was proposed by Pepe & Anderson (1994), albeit in the context of a longitudinal study with a mean model $E(Y_{it}|X_{it})$. Here, the subscript i represents an individual, and t denotes time. Pepe & Anderson (1994) argued that if the implicit assumption $E(Y_{it}|X_{it}) = E(Y_{it}|X_{it}, X_{it'}, t' \neq t)$ does not hold, then the independence working covariance becomes the sole valid option. Pan et al. (2000) derived the bias for GEE with a dependent working covariance for a specific linear model. This bias formula offers valuable insights into understanding this seemingly counter-intuitive conclusion. In our paper, we explore the scenario $E(Y_{it}|Y_{it'}, t' \neq t, X_i)$ and Y_{it}'s are binary and reach a similar conclusion to Pepe & Anderson (1994).

The rest of this article is arranged as follows. In Section 2, we rewrite the QEBD in a quadratic form and derive its score function and Fisher's information matrix. In Section 2, we consider estimating equations based on partial conditional distributions to alleviate the computation burden and prove that the independence working correlation is the only valid one to yield consistent estimates among other popular working correlations. We also define a QELR in Section 2.3, followed by simulation studies in Section 3 and real data analyses in Section 4. Conclusions are drawn in Section 5.

2 Quadratic Exponential Logistic Regressions

2.1 Quadratic Exponential Binary Distributions

We first fix the notation. Following the convention, we denote capital letters as matrics, e.g., A and B; bold-faced letters as vectors, e.g., $\mathbf{h}$ and $\mathbf{y}$; bold-faced capital letter as a vector consisting of random variables, say $\mathbf{Y}$. For an $n \times m$ matrix B, let $[B]_{ij}$ be the (i,j)th element of B. Denote $A \otimes B$ as the Kronecker product of matrices A and B. When $A \in \mathbb{R}^{2\times 3}$ and B is defined above,

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & a_{13}B \\ a_{21}B & a_{22}B & a_{23}B \end{bmatrix} \in \mathbb{R}^{2n\times 3m}$$

where $a_{ij} = [A]_{ij}$. Moreover, let $vec(\cdot)$ be an operator that vectorizes its argument into a vector, e.g., $vec(A) = (a_{11}, a_{21}, a_{12}, a_{22}, a_{13}, a_{23})^\top$. When A, B and C are three conformable matrices, $vec(ABC) = \left(C^\top \otimes A\right) vec\,(B)$. This identity is frequently applied later. Last, let $\mathbf{Y}_{[j]} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_m)$. Then define $\mathbf{y}_{[j]} \in \mathbb{R}^{m-1}$ as the corresponding realization of $\mathbf{Y}_{[j]}$, and $\mathbf{y}^x_{[j]} \in \mathbb{R}^m$ as the vector of $\mathbf{y}$ but substitutes x to the jth element of $\mathbf{y}$. For example, if $\mathbf{y} = (y_1, y_2, y_3)$ then $\mathbf{y}_{[3]} = (y_1, y_2)$ and $\mathbf{y}^1_{[3]} = (y_1, y_2, 1)$. Throughout this paper, let $\mathbf{e}_j$ be the jth column of the m-dimensional identity matrix for $j = 1, \ldots, m$.

Suppose that the data consist of n independent m-tuple of binary responses, say $(Y_{k1}, \ldots, Y_{km})$ where $k = 1, \ldots, n$. Each of Y_{kj}'s takes values 0 or 1. In this section, we first suppress the subject subscript k for convenience and will re-express it when considering regression models. Following Cox (1972), Zhao & Prentice (1990), and Cox & Wermuth (1994), the QEBD has the form

$$\Pr(Y_1 = y_1, \ldots, Y_m = y_m) = \exp\left\{\sum_{j=1}^{m} y_j\beta_j + \sum_{i<j} \theta_{ij} y_i y_j - \Lambda\right\} \tag{1}$$

where Λ is the normalizing constant which consists of 2^m terms for all possible configura-

tions of $(y_1, \ldots, y_m)$. This model is also the asymmetric Ising model defined in Anandkumar et al. (2012). Let $\mathbf{y} = (y_1, \ldots, y_m)^\top$ and Θ be an $m \times m$ symmetric matrix such that $[\Theta]_{jj} = \beta_j$ and $[\Theta]_{ij} = \theta_{ij}/2 = [\Theta]_{ji}$ for $i < j$. We collect all unique parameters in Θ into $\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \ldots, \theta_{1m}, \theta_{22}, \theta_{23}, \ldots, \theta_{mm})^\top \in \mathbb{R}^{m_2}$ where $m_2 = (m+1)m/2$. Then, because $y_i^2 = y_i$,

$$f(\mathbf{y}) = \Pr(Y_1 = y_1, \ldots, Y_m = y_m) = \exp\left\{\mathbf{y}^\top \Theta \mathbf{y} - \Lambda\right\}$$

which has exactly the same form as the kernel of a multivariate normal distribution. Unfortunately, this distribution is not closed under taking arbitrary marginals (Joe & Liu 1996). Worse, solving θ_{ij}'s requires evaluating Λ repeatedly and hence causes a massive computation burden.

2.2 Fully Conditional Distributions

Collecting the partial conditional probabilities is the first step for constructing estimating equations. The conditional probability of one variable conditional on the rest has the probability function $\Pr(Y_j = y_j | \mathbf{Y}_{[j]} = \mathbf{y}_{[j]}) = \pi_j^{y_j}(1-\pi_j)^{1-y_j}$ with the logistic regression form, for $j = 1, \ldots, n$,

$$\pi_j = \frac{\exp\left\{\beta_j + \sum_{i \neq j} [\tilde{\Theta}]_{ij} y_i\right\}}{1 + \exp\left\{\beta_j + \sum_{i \neq j} [\tilde{\Theta}]_{ij} y_i\right\}} = \frac{\exp\left\{\mathbf{e}_j^\top \tilde{\Theta} \mathbf{y}_{[j]}^1\right\}}{1 + \exp\left\{\mathbf{e}_j^\top \tilde{\Theta} \mathbf{y}_{[j]}^1\right\}} \tag{2}$$

where $[\tilde{\Theta}]_{jj} = \beta_j$ and $[\tilde{\Theta}]_{ij} = 2[\Theta]_{ij} = \theta_{ij}$. When $i < j$, the parameter $\exp\{[\tilde{\Theta}]_{ij}\} = \exp\{\theta_{ij}\}$ is therefore the odds ratio of Y_j conditional on Y_i given the rest responses, Y_l's, $l \neq i, j$, remain unchanged.

Connolly & Liang (1988) suggested summing the above score functions to form estimating equations. At first glance, Connolly & Liang (1988) applied GEE approach with independence working covariance. However, general applications of the GEE approach aggregate the score functions of marginal distributions, not conditional distributions. We argue

that this discrepancy does not alter the desired large sample properties in the view of M-estimation, e.g., Stefanski & Boos (2002), because $E(\mathbf{y}^1_{[j]}(y_j - \pi_j)) = \mathbf{0}$ for all $j = 1, \ldots, m$, by the double expectation rule. Moreover, inserting a working correlation, say $V \in \mathbb{R}^{m \times m}$, is reasonable to account for potential correlation among these conditional variables. The estimation functions for $\boldsymbol{\theta}$ can be

$$\varphi^F(\boldsymbol{\theta}; V) = G\left[\mathbf{y}^1_{[1]} \otimes \mathbf{e}_1, \ldots, \mathbf{y}^1_{[m]} \otimes \mathbf{e}_m\right] V(\mathbf{y} - \boldsymbol{\pi}) \tag{3}$$

where $\mathbf{e}_j$ is the jth column of the $m \times m$ identity matrix and $G \in \mathbb{R}^{m(m+1)/2 \times m^2}$ is defined in Appendix A. The superscript F stands for the "fully"-conditional model. Surprisingly, under an arbitrary working correlation, the expectation of (3) is not necessarily vanishing as the sample size goes to infinity. Hence solving $\varphi^F(\boldsymbol{\theta}; V) = \mathbf{0}$ may yield biased estimations except for some carefully chosen working correlation V. As summarized in Theorem 2.1, the independence working covariance is the only valid working covariance that always yields consistent estimates. The proof is deferred to Appendix A.

Theorem 2.1. *Let $\mathbf{Y} \in \mathbb{R}^m$ follow the QEBD as in (1) with estimating functions (3). Then $E_{\mathbf{Y}}(\varphi^F(\boldsymbol{\theta}; V)) = \mathbf{0}$ if and only if V is a diagonal matrix.*

In brief, we proved that, instead of directly maximizing the QEBD likelihood (1), solving the estimating equations of the fully conditional distribution yields consistent estimates, too, when the independence working covariance is enforced. Next, since the fully conditional distribution has the form of logistic regression, we may introduce covariates to the model to broaden the applications of the QEBD. The remaining question is whether an arbitrary set of fully conditional distributions results from a unique QEBD. The answer is positive unless the compatibility condition is satisfied, Arnold & Press (1989).

2.3 Regression for Quadratic Exponential Binary Responses

For clarity, we name β_j as the main effect and $[\tilde{\Theta}]_{ij}$ as the interaction effect. Constructing models for the main effect is relatively simple. Suppose $\mathbf{x}_j$ is a vector of predictors affecting the individual effect. Then, we can define $\beta_j = \boldsymbol{\beta}^\top \mathbf{x}_j$. Modeling the main effects of the QELR in this way has been proposed in the literature, such as Connolly & Liang (1988), Zhao & Prentice (1990), Joe & Liu (1996), and many others.

Next, we deal with the interaction effect, $[\tilde{\Theta}]_{ij}$'s. As demonstrated in Arnold & Press (1989), arbitrarily modifying the fully conditional distributions like (2) may not yield a unique joint distribution like (1). The compatibility condition is required to ensure the existence and uniqueness of the joint distribution. The compatibility condition for the model (2) are $[\tilde{\Theta}]_{ij} = [\tilde{\Theta}]_{ji}$ for all $i \neq j$, Joe & Liu (1996). Consequently, when having extra information about the interaction effects, we may define $[\tilde{\Theta}]_{ij} = \gamma w_{ij}$ where w_{ij} is a predictor representing the cause of the interaction between the ith variable and the jth variable, and γ is the corresponding regression coefficient. Additionally, we need to enforce that $\gamma w_{ij} = \gamma w_{ji}$ to satisfy the compatible condition.

When there are L characteristics that potentially describes the interactions among variable y_j's, to meet the compatibility condition, we define $w_{ij}^\ell = \text{dist}(\mathbf{u}_i^\ell, \mathbf{u}_j^\ell)$ where $\mathbf{u}_j^\ell$ is the observed vector of the ℓth characteristic about the jth variable, and a distance function $\text{dist}(\mathbf{u}_i^\ell, \mathbf{u}_j^\ell) \geq 0$ which returns the distance between two vectors where $\text{dist}(\mathbf{u}_i^\ell, \mathbf{u}_j^\ell) = 0$ if and only if $\mathbf{u}_i^\ell = \mathbf{u}_j^\ell$ and the distance function has to be symmetric, $\text{dist}(\mathbf{u}_i^\ell, \mathbf{u}_j^\ell) = \text{dist}(\mathbf{u}_j^\ell, \mathbf{u}_i^\ell)$. The proposed QELR is, therefore, for $j = 1, \ldots, m$, having the fully conditional log-density function

$$\text{logit}\left(\Pr(Y_j = 1 | \mathbf{Y}_{[j]} = \mathbf{y}_{[j]})\right) = \boldsymbol{\beta}^\top \mathbf{x}_j + \sum_{i \neq j} \sum_{\ell=1}^{L} \gamma^\ell w_{ij}^\ell y_i = \boldsymbol{\beta}^\top \mathbf{x}_j + \boldsymbol{\gamma}^\top W_j \mathbf{y}_{[j]}^0 \qquad (4)$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_L)^\top$ and

$$W_j = \begin{bmatrix} w_{1j}^1 & w_{2j}^1 & \ldots & w_{mj}^1 \\ w_{1j}^2 & w_{2j}^2 & \ldots & w_{mj}^2 \\ \vdots & \vdots & \ddots & \vdots \\ w_{1j}^L & w_{2j}^L & \ldots & w_{mj}^L \end{bmatrix} \in \mathbb{R}^{L \times m}.$$

The most simple version of (4) is setting a common interaction effect. That is,

$$\text{logit}\left(\Pr(Y_j = 1 | \mathbf{Y}_{[j]} = \mathbf{y}_{[j]})\right) = \beta_j + \sum_{i \neq j} \theta y_i = \beta_j + \theta \left(\sum_{i=1}^{m} y_i - y_j \right) \tag{5}$$

for $j = 1, \ldots, n$. In contrast to the interaction effects in (4), this model has $\gamma = \theta \quad \text{and} \quad W_j = [1, \ldots, 1] \in \mathbb{R}^{1 \times m}$. This equation implies the symmetry among variable in $\mathbf{y}_{[i]}$ and thus $\sum_{i=1}^{m} y_i$ is a sufficient statistic for θ, Connolly & Liang (1988). Qu et al. (1987) and Connolly & Liang (1988) considered a more general model, $\beta_j + F_{\boldsymbol{\alpha}}(\sum_{i=1}^{m} y_i - y_j)$ where $F_{\boldsymbol{\alpha}}$ is a known function with unknown parameters $\boldsymbol{\alpha}$.

With the compatible condition, defining all fully conditional distributions like (4) results in a unique joint model. Solving the joint model likelihood is challenging. The difficulty arises due to evaluating the normalizing term, which consists of 2^m terms per observation. According to the previous discussion and our simulation results listed in Simulation Studies, we suggest using the GEE approach. According to (4), the resulting estimating functions are

$$\varphi^R(\boldsymbol{\theta}; V) = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \ldots & \mathbf{x}_m \\ W_1 \mathbf{y}_{[1]}^0 & W_2 \mathbf{y}_{[2]}^0 & & W_m \mathbf{y}_{[m]}^0 \end{bmatrix} V(\mathbf{y} - \boldsymbol{\pi}) \tag{6}$$

where V is a working covariance matrix and the superscript R emphasizes that the estimating functions contain regressors. Note that the upper half of the estimating functions in (6) has the form $[\mathbf{x}_1, \ldots, \mathbf{x}_m]V(\mathbf{y} - \boldsymbol{\pi})$ which has mean zero. On the other hand, the lower half of the estimating functions have a complicated form of y_i's, which may not yield

zero means. Hence, the consistency of the estimation is questionable. Again, arbitrary choices of the working covariance are not guaranteed to end up with consistent estimates. Similar to Theorem 2.1, the independence covariance is a valid choice for consistency. We summarize our conclusion in Theorem 2.2, and its proof is deferred to Appendix B.

Theorem 2.2. *The fully conditional functions defined in (4) result in* $E_{\mathbf{Y}}(\varphi^R(\boldsymbol{\theta}; V)) = \mathbf{0}$*, provided* V *is diagonal.*

3 Simulations

In the following, we apply four estimation approaches to data analysis for different purposes. The first approach is the maximum likelihood estimate, MLE. We maximize the likelihood function (1) directly. It is time-consuming for mild m, say $m = 15$. The next two approaches are the GEE approach with two different working covariances. We abbreviate GEE with independence working covariance as GEE-IND and GEE with exchangeable working covariance as GEE-EXC. These two approaches were calculated by the *geepack* package, Halekoh et al. (2006), prepared in the R language, R Core Team (2021). Last, the generalized linear model (GLM) approach is applied by ignoring the potential correlations among conditional variables. According to the large sample property of these approaches, we have the following foresight. The estimates of GLM and GEE-IND are identical, but their standard error estimates are different. When the data follow a QEBD, the standard error estimates of GLM are incorrect, more or less. For GEE-EXC, since the resulting estimators are inconsistent by Theorem 2.1, the standard error estimates are meaningless. Finally, although solving MLE is time-consuming, we apply MLE to conclude that GEE-IND and MLE are comparable under applicable scenarios. R codes for simulation studies are available at `https://github.com/jonong03/QELR/`.

3.1 Simulation I: Quadratic Exponential Distributions

The first simulation study evaluated the estimation quality and computational efficiency of the MLE, GLM, and GEE approaches. The data followed the QEBD (1) with m binary responses based on a given set of true parameters with a fixed sample size (n=300). For each scenario, the parameters were estimated by MLE, GLM, and GEE-IND approaches. The GEE-EXC was ignored because it always diverged in this simulation setup. Five hundred datasets were simulated for data analysis. The simulation results were summarized in Table 1. The estimation of bias (mean of parameter estimates subtracting the true value) and the average standard error (mean of standard error estimates) were reported. Both MLE and GEE-IND yielded negligible biases and almost identical standard error estimates on average. However, although the GLM approach yields the same estimates as the GEE-IND, its standard error estimates were far too small. We conclude that the MLE and GEE-IND are numerically comparable.

To understand the computing costs for the aforementioned methods, we fixed the sample size as $n = 300$ and increased the number of binary variables m. The average computing times of the MLE method for $m = 5$, 10, 12, are 0.652, 26.346, and 196.759 seconds, respectively, whereas the average computing times of the GEE-IND method for $m = 5$, 10, 12, are 0.064, 0.213, and 0.320 seconds, respectively. The MLE method required a much longer execution time than the GEE-IND and could pose computational challenges, especially as the number of parameters increases (e.g., m=15 or larger). In contrast, the advocated GEE-IND approach offers a feasible solution to this computational issue.

Table 1: Parameter Estimations for the QEBD (n=300, m=5, replicates=500)

		Emp.	MLE		GLM		GEE-IND	
Variable	Truth	S.D.	Bias	R.E.	Bias	R.E.	Bias	R.E.
β_1	-1.500	0.439	-0.063	1.045	-0.064	0.766	-0.064	1.046
β_2	-0.750	0.381	-0.015	0.989	-0.015	0.737	-0.015	0.988
β_3	0.000	0.333	0.000	0.996	0.000	0.737	0.000	0.995
β_4	0.750	0.328	0.018	0.987	0.018	0.746	0.018	0.985
β_5	1.500	0.318	0.017	1.009	0.017	0.777	0.017	1.004
θ_{12}	-0.400	0.385	-0.025	0.962	-0.024	0.676	-0.024	0.961
θ_{13}	0.400	0.317	-0.015	0.932	-0.015	0.655	-0.015	0.932
θ_{14}	-0.400	0.309	-0.012	0.940	-0.012	0.661	-0.012	0.941
θ_{15}	0.400	0.400	0.054	1.043	0.054	0.734	0.054	1.043
θ_{23}	-0.400	0.280	-0.014	1.004	-0.014	0.705	-0.014	1.004
θ_{24}	0.400	0.291	0.011	0.992	0.011	0.696	0.011	0.992
θ_{25}	-0.400	0.343	-0.009	0.971	-0.009	0.682	-0.009	0.970
θ_{34}	-0.400	0.249	-0.027	0.980	-0.027	0.688	-0.027	0.979
θ_{35}	0.400	0.316	0.027	0.976	0.027	0.686	0.027	0.974
θ_{45}	-0.400	0.323	0.005	0.986	0.005	0.692	0.005	0.984

Emp. S.D.: the shorthand of the empirical standard deviation, which is the sample standard deviation of the 500 MLEs. R.E.: relative efficiency which is the mean standard error estimate over the 500 replicates divided by the Emp. S.D..

3.2 Simulation II: Quadratic Exponential Logistic Regressions

Next, we present a simulation study to assess the performance of our proposed QELRs with common interaction (5) and linear interaction (4). We simulated 500 datasets, each consisting of n individuals and $m = 15$ correlated binary responses. The sample sizes n are 100, 300, and 500. In particular, the fully conditional distribution for the common interaction model is

$$\text{logit}\left(\Pr(Y_j = 1|\mathbf{Y}_{[j]} = \mathbf{y}_{[j]})\right) = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \gamma\left(\sum_{i=1}^{m} y_i - y_j\right)$$

and the fully conditional distribution for the linear interaction model is

$$\text{logit}\left(\Pr(Y_j = 1|\mathbf{Y}_{[j]} = \mathbf{y}_{[j]})\right) = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \sum_{i=1, i\neq j}^{m} (\gamma_1 w_{ij}^1 + \gamma_2 w_{ij}^2) y_i.$$

Note that w_{ij}^{ℓ} is a similarity measure of the ith and the jth variables. In the simulation, we generated a pair of covariates for each variable, say (u_j^1, u_j^2), $j = 1, \ldots, m$. Then $w_{ij}^{\ell} = I(u_i^{\ell} = u_j^{\ell})$, for $i, j = 1, \ldots, m$ and $\ell = 1, 2$. Here, u_i^{ℓ}'s were sampled from the set $\{1, 2, 3\}$ uniformly so that $\Pr(w_{ij}^{\ell} = 1) = \Pr(u_i^{\ell} = u_j^{\ell}) = 1/3$. The compatibility condition of Joe & Liu (1996) is satisfied by doing so.

Analogous to the previous study, we tried the GLM, GEE-IND, and GEE-EXC approaches. The simulation results from each scenario in Tables 2 and 3 revealed that GLM and GEE-IND exhibited negligible biases. According to the relative efficiency of the GLM columns in Tables 2 and 3, we conclude that the standard error estimates of GLM were too small. Hence, the GLM approach may result in too many significant (from zero) covariates. On the other hand, GEE-EXC demonstrated prominent biases for the common interaction case but inconspicuous biases for the linear interaction case. Additionally, the GEE-EXC under the common interaction model yielded a 57% divergence rate on average but 0% for the linear interaction model. The inconsistency is due to the sparseness of

the correlation structures. All correlation elements of the common interaction model take value γ, whereas only one-third of w_{ij}^{ℓ} of the linear interaction model take nonzero values on average. Through this simulation, we emphasize that GEE-IND guarantees estimation consistency, but other working covariances may yield biased estimations. The degree of bias may depend on the sparsity of the correlation structure.

4 Case Studies

4.1 English Elite Course Survey

To bolster students' international mobility and elevate their English proficiency and confidence for future study abroad opportunities, a vocational university in Taiwan took a proactive step by introducing the English Elite Course. This specialized course was tailored to prepare first-year students for the TOEIC (Test of English for International Communication) examination supported by the Bilingual Education Project. A simple survey was conducted to gauge the effectiveness of the English Elite Course and identify areas for potential teaching enhancements. The survey consisted of seven carefully crafted questions designed to assess the student's learning status to the current course learning goals. The primary objectives of the survey were to gain insights into the student's language learning progress and to solicit valuable feedback that could be used to refine the course content and teaching methodologies. By incorporating student input, the university aimed to strengthen language skills, boost English proficiency, and instill greater confidence in students, thereby empowering them for success in their future endeavors studying abroad. These questions (with shorthand in the parentheses) are:

1. Participating in this course has helped me improve my English (Improve).

Table 2: Estimation Results of the QELR with Common Interaction Effects (m=15)

Sample			Emp.	GLM		GEE-IND		GEE-EXC*	
Size	Variable	Truth	S.D.	Bias	R.E.	Bias	R.E.	Bias	R.E.
n=100	β_0	-2.400	0.594	0.042	0.627	0.042	0.975	0.918	0.976
	β_1	-2.000	0.220	-0.051	0.978	-0.051	0.982	0.161	0.886
	β_2	-2.600	0.265	-0.068	0.948	-0.068	0.957	0.216	0.867
	γ	-1.400	0.298	-0.082	0.555	-0.082	0.937	-0.438	0.999
n=300	β_0	-2.400	0.333	0.014	0.623	0.014	0.965	-0.863	0.943
	β_1	-2.000	0.119	-0.015	1.019	-0.015	1.041	0.638	0.929
	β_2	-2.600	0.142	-0.020	0.992	-0.020	1.018	-0.301	0.924
	γ	-1.400	0.167	-0.023	0.546	-0.023	0.938	0.969	1.026
n=500	β_0	-2.400	0.247	0.003	0.649	0.003	1.000	-0.127	0.991
	β_1	-2.000	0.090	-0.011	1.034	-0.011	1.059	-1.248	0.941
	β_2	-2.600	0.104	-0.011	1.041	-0.011	1.074	0.127	0.969
	γ	-1.400	0.125	-0.013	0.560	-0.013	0.963	0.371	1.061

Emp. S.D.: the shorthand of the empirical standard deviation, the sample standard deviation of the 500 MLEs.

R.E.: relative efficiency is the mean standard error estimate over the 500 replicates divided by the Emp. S.D..

*: the divergence rates for sample sizes 100, 300, and 500 are 53%, 58%, and 60%, respectively.

Table 3: Estimation Results of the Quadartic Exponential Logistic Regression with Linear Interaction Effects (m=15)

Sample			Emp.	GLM		GEE-IND		GEE-EXC	
Size	Variable	Truth	S.D.	Bias	R.E.	Bias	R.E.	Bias	R.E.
n=100	β_0	-2.400	0.242	-0.008	0.882	-0.008	1.000	0.127	0.997
	β_1	-2.000	0.145	-0.035	1.083	-0.035	1.084	-0.018	1.076
	β_2	-2.600	0.178	-0.031	1.032	-0.031	1.037	-0.009	1.032
	γ_1	-1.400	0.238	-0.045	0.761	-0.045	0.976	-0.123	0.990
	γ_2	-0.500	0.193	-0.023	0.807	-0.023	1.011	-0.095	1.028
n=300	β_0	-2.400	0.142	0.002	0.860	0.002	0.981	0.135	0.976
	β_1	-2.000	0.087	-0.012	1.030	-0.012	1.038	0.004	1.031
	β_2	-2.600	0.104	-0.011	1.009	-0.011	1.018	0.009	1.012
	γ_1	-1.400	0.136	-0.018	0.757	-0.018	0.989	-0.097	1.005
	γ_2	-0.500	0.111	-0.008	0.802	-0.008	1.030	-0.082	1.049
n=500	β_0	-2.400	0.113	0.003	0.833	0.003	0.953	0.135	0.948
	β_1	-2.000	0.067	-0.008	1.026	-0.008	1.033	0.008	1.027
	β_2	-2.600	0.082	-0.010	0.992	-0.010	1.003	0.010	0.997
	γ_1	-1.400	0.108	-0.009	0.738	-0.009	0.973	-0.088	0.990
	γ_2	-0.500	0.089	-0.011	0.771	-0.011	0.988	-0.084	1.007

Emp. S.D.: the shorthand of the empirical standard deviation, the sample standard deviation of the 500 MLEs.

R.E.: relative efficiency is the mean standard error estimate over the 500 replicates divided by the Emp. S.D..

2. After participating in this course, I would like to challenge myself to achieve a higher score (Challenge).

3. Participating in this course has enhanced my motivation and confidence in learning English (Confidence).

4. Overall, I have acquired significant knowledge from this course. (Knowledge).

5. I feel that the TOEIC certification will greatly benefit me in the future (Benefit).

6. I hope the school can offer more English enhancement classes for TOEIC exams (More Courses).

7. I hope to have the opportunity to be an exchange student or study/intern abroad during my four years of university life (Exchange).

Students' responses were recorded on a five-Liker scale, valued 1 as very disagree and 5 as strongly agree. We re-coded them into 1 as the response is greater or equal to 3 and 0 otherwise.

Data analysis results are shown in Table 4. We first consider the interpretation of parameters. Since the fully conditional distribution of the Improve variable is

$$\begin{aligned}
&\log\left(\frac{\Pr(\text{Improve}=1)}{\Pr(\text{Improve}=0)}\right) \\
=& -1.641 - 0.491 \times I(\text{Challenge}=1) + 2.137 \times I(\text{Confidence}=1) \\
&+ 2.569 \times I(\text{Knowledge}=1) + 0.320 \times I(\text{Benefit}=1) \\
&- 0.059 \times I(\text{More Courses}=1) - 0.047 \times I(\text{Exchange}=1).
\end{aligned}$$

The interpretation of regression coefficients is thus the same as the logistic regression. As the introduction mentions, the QEBD (asymmetric Ising model) is commonly employed in network analysis. In Figure 1, we visualize the estimates listed in Table 4 in a network

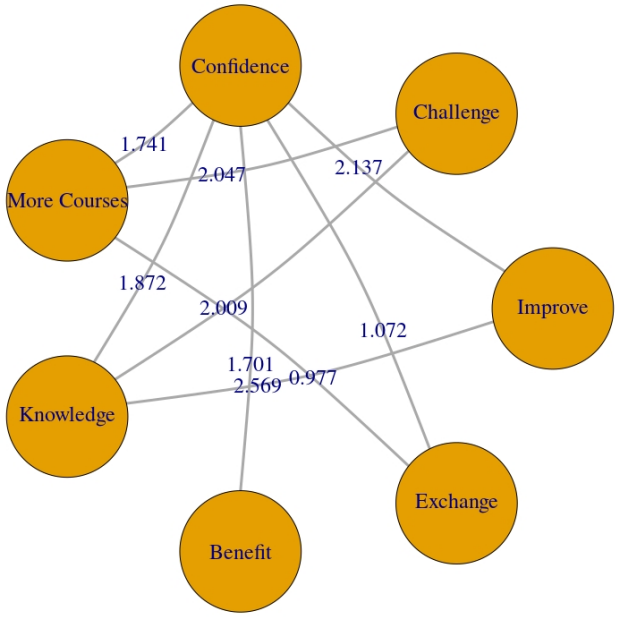

Figure 1: Network Summery of the Advance English Course Survey. The presented edges are statistically different from zero.

(or undirected graph). Here, the questions are represented as vertices and the estimates of θ_{ij} that are significantly different from 0 are depicted as edges. By examining this network representation, several key insights can be derived. First, it becomes evident that increasing students' tendency to provide higher ratings for Confidence should be a primary learning goal of the study plan. This is supported by the fact that Confidence is linked to five other questions (Improve, Knowledge, More Courses, Benefit, and Exchange) with positive regression coefficients: 2.137, 1.872, 1.741, 1.701, and 1.072, respectively. Second, to encourage students to participate in exchange programs, additional strategies should be directed toward higher ratings in Confidence, More Courses, and Challenges. By visualizing the information in this manner, we gain a clearer understanding of the interconnections between different questions and can derive actionable insights to effectively inform the teaching and learning process.

Table 4: Analysis of the English Elite Course Survey ($n = 193$)

Variables	Estimate	S.E.	z-value	p-value	Prop.
Main Effects					
Improve	**-1.641**	0.395	-4.154	0.000	0.606
Challenge	**-3.293**	0.498	-6.618	0.000	0.503
Confidence	**-5.915**	0.962	-6.146	0.000	0.477
Knowledge	**-2.837**	0.526	-5.396	0.000	0.591
Benefit	**-0.887**	0.294	-3.017	0.003	0.689
More Courses	**-2.552**	0.456	-5.592	0.000	0.461
Exchange	**-1.669**	0.365	-4.574	0.000	0.508
Interaction Effects					
Improve-Challenge	-0.491	0.711	-0.691	0.490	0.409
Improve-Confidence	**2.137**	0.644	3.320	0.001	0.440
Improve-Knowledge	**2.569**	0.536	4.795	0.000	0.523
Improve-Benefit	0.320	0.521	0.614	0.539	0.513
Improve-More Courses	-0.059	0.740	-0.080	0.936	0.373
Improve-Exchange	-0.047	0.474	-0.098	0.922	0.383
Challenge-Confidence	0.635	0.600	1.059	0.290	0.389
Challenge-Knowledge	**2.009**	0.657	3.060	0.002	0.440
Challenge-Benefit	0.930	0.574	1.620	0.105	0.456
Challenge-More Courses	**2.047**	0.469	4.366	0.000	0.389
Challenge-Exchange	0.847	0.468	1.811	0.070	0.373
Confidence-Knowledge	**1.872**	0.584	3.206	0.001	0.446
Confidence-Benefit	**1.701**	0.720	2.363	0.018	0.451
Confidence-More Courses	**1.741**	0.585	2.976	0.003	0.368
Confidence-Exchange	**1.072**	0.523	2.052	0.040	0.363
Knowledge-Benefit	0.647	0.557	1.162	0.245	0.518
Knowledge-More Courses	-0.184	0.659	-0.280	0.780	0.383
Knowledge-Exchange	-0.281	0.536	-0.524	0.600	0.389
Benefit-More Courses	0.060	0.547	0.109	0.913	0.409
Benefit-Exchange	0.755	0.463	1.630	0.103	0.440
More Courses-Exchange	**0.977**	0.428	2.284	0.022	0.352

Prop.: proportions of, for main effects, the variable being one and, for interactions effects, both variables being one.

4.2 Influence Among Justices in Taiwan Constitutional Court Opinions

In appellate court proceedings, the ultimate decision emerges from a series of choices that each judge makes at various points in a case. Judges don't work in isolation; instead, they engage in a collaborative process with their colleagues to develop an opinion that represents the court's collective stance. This interplay among judges significantly shapes the final decision. Those who align with the case outcome but diverge in legal reasoning have the option to join or author a concurrence. Conversely, judges who find themselves at odds with both the case outcome and the majority's legal rationale have the liberty to affiliate with or compose a dissent.

The field is replete with research exploring the determinants of judges' votes and non-consensual opinions. A considerable segment of this literature examines collective influences, investigating how factors such as judges' personal characteristics (e.g., ideologies, gender, prior work experience, education, as cited in Cross & Tiller (1998)), institutional norms, and leadership (Walker et al. (1988), Haynie (1992)), or a combination of both personal and institutional elements over time (using models like logit regression and the Partial Proportional Odds model, as per Ward et al. (2023); or Autoregressive Models, as used by Hall & Windett (2016)) sway the panel's consensus. In contrast, other studies (e.g., Revesz (1997); Farhang & Wawro (2004); Peresie (2005); Boyd et al. (2010)) focus on individual-level analyses, investigating how judges' personal attributes shape their voting behavior.

Divergent approaches have been adopted to explore the influence of peer dynamics on judges' votes and opinions. While a significant portion of the literature utilizes logit or probit regression models (e.g., Wahlbeck et al. (1999)), Zorn (2001) stands out as a

trailblazer, having employed the GEE approach to tackle the issues of dependencies over time and among justices in judicial decision-making. An alternative stream of research takes into account peer effects by scrutinizing peers' votes through the lens of simultaneous equations models (Fischman (2015), Holden et al. (2021)).

Our study seeks to understand how justices' social networks, encompassing shared educational or professional backgrounds, affect their inclination to align with each other's opinions. We hypothesize a tight interweaving of opinion writing and interdependent voting. Justices first analyze the issue at hand and the case outcome, guided by the collective votes and predominant reasoning. They then weigh in their colleagues' votes and rationales before aligning with a concurring or dissenting stance. These issues and case outcomes are considered critical information that justices draw upon while collaboratively crafting a non-consensual opinion. As their tenure progresses, justices naturally grow more acquainted with each other, further enriching this collaborative process.

As a confirmative study, we employ QELR to scrutinize the Taiwan Constitutional Court (TCC) dataset, focusing specifically on the influence exerted among justices during the opinion writing process. Of note is the increasing trend in opinion writing frequency among TCC justices in recent years. Our sample is drawn from the October 2016 to September 2019 term. By confining our analysis to a relatively brief period, we strive to segregate the effects of personal characteristics on inter-justice interactions from substantial shifts in TCC membership and related jurisprudence over time.

Our dataset includes 344 opinions from 2016 to 2019. We coded several variables for each opinion: the contributing justices (represented by a 15-level categorical variable for the 15 justices), the issue at stake (a 3-level categorical variable incorporating constitutional rights, constitutional institutions, and legal rights), the case outcome (a 3-level categorical

variable denoting constitutional ruling and unconstitutional ruling), and the justices' tenure length in years. These variables are integrated into the main effect model. For interaction effects, we accounted for the justices' educational backgrounds (distinguishing whether a pair of justices both received foreign degrees from countries practicing either the common law or civil law system or whether neither has a foreign degree) and their prior professional experiences (indicating whether a pair of justices shared the same previous occupation, divided into academic or legal jobs). The results of the data analysis are presented in Table 5. From this table, it appears that the issue at hand impacts the main effect, but the case outcome does not. Similarly, justices' shared educational backgrounds influence the interaction effects, while their previous occupations do not.

5 Conclusion

This work suggests a GEE approach for efficient and consistent estimation of the QEBD and QELR. We proved that the independence working covariance is the only one to guarantee the estimation consistency. Sometimes, the non-independence working covariance causes divergence of GEE and, sometimes, causes mild biases only. It depends on the underlying models, which are often unknown to us. The QEBD considers one-way main effects and two-way interaction effects only to represent the model as a network. The English elite course survey is an example. Furthermore, the proposed QELR can be used to confirm the hypothetical correlation structures among mutually affected individuals, as demonstrated in the non-consensual opinion writing examples.

The fully conditional model has exactly the form of a logistic regression; hence, the model inherits the pros and cons of the logistic regression. Firth (1993) pointed out that some true parameter values do not exist when the data is separable. Adding certain penalty

Table 5: Grand Justice Data Analysis Using QELR

	Estimate	S.E.	z-value	p-value
Main Effects				
Judge ID				
GJ1	**-2.345**	0.349	-6.719	0.000
GJ2	**-2.577**	0.355	-7.254	0.000
GJ3	**-1.626**	0.327	-4.969	0.000
GJ4	**-1.940**	0.372	-5.213	0.000
GJ5	**-1.773**	0.355	-4.996	0.000
GJ6	**-1.720**	0.342	-5.034	0.000
GJ7	**-3.422**	0.443	-7.729	0.000
GJ8	**-3.786**	0.478	-7.913	0.000
GJ9	**-1.627**	0.286	-5.692	0.000
GJ10	**-1.573**	0.306	-5.138	0.000
GJ11	**-2.383**	0.397	-6.002	0.000
GJ12	**-2.332**	0.318	-7.343	0.000
GJ13	**-1.663**	0.310	-5.368	0.000
GJ14	**-2.601**	0.371	-7.011	0.000
GJ15	**-1.537**	0.320	-4.811	0.000
Issue				
Constitutional Rights	-0.357	0.233	-1.533	0.063
Constitutional Institutions	**-0.408**	0.247	-1.652	0.049
Case Outcome				
Constitutional Ruling	0.270	0.197	1.371	0.085
Unconstitutional Ruling	0.293	0.204	1.439	0.075
Time (t)				
t	-1.078	2.354	-0.458	0.323
t^2	3.741	6.203	0.603	0.273
t^3	-2.753	3.878	-0.710	0.239
Interaction Effects				
Prior Occupation	0.065	0.254	0.258	0.398
Foreign Education	**-1.288**	0.247	-5.224	0.000

terms can be a solution. However, this is beyond our scope, and we defer it to future studies.

Using accumulated fully conditional models raises another computing difficulty. To illustrate, consider a dataset comprising n samples, each associated with m binary responses, and consequently, the design matrix for the proposed regression comprises $n \times m$ rows. The sheer size of this design matrix can potentially lead to computational issues. Enea (2009) has proposed strategies for handling large datasets in generalized linear models, and it would be valuable to invest efforts in adapting and integrating these strategies into GEE computations.

References

Anandkumar, A., Tan, V., Huang, F. & Willsky, A. (2012), 'High-dimensional structure estimation in Ising models: local separation criterion', *The Annals of Statistics* **40**, 1346–1375.

Anderson, C., Wasserman, S. & Crouch, B. (1999), 'A $p^\star$ primer: logit models for social networks', *Social Networks* **21**, 37–66.

Arnold, B. & Press, S. (1989), 'Compatible conditional distributions', *Journal of the American Statistical Association* **84**, 152–156.

Boyd, C. L., Epstein, L. & Martin, A. D. (2010), 'Untangling the causal effects of sex on judging', *American Journal of Political Science* **54**(2), 389–411.

Breslow, N. E. & Clayton, D. G. (1993), 'Approximate inference in generalized linear mixed models', *Journal of the American Statistical Association* **88**(421), 9–25.

Connolly, M. & Liang, K. (1988), 'Conditional logistic regression models for correlated binary data', *Biometrika* **75**, 501–506.

Cox, D. (1972), 'The analysis of multivariate binary data', *Journal of the Royal Statistical Society. Series C* **21**, 113–120.

Cox, D. R. & Wermuth, N. (1994), 'A note on the quadratic exponential binary distribution', *Biometrika* **81**, 403–408.

Crespi, C., Wong, W. & Mishra, S. (2009), 'Using second-order generalized estimating equations to model heterogeneous intraclass correlation in cluster randomized trials', *Statistics in Medicine* **28**, 814–827.

Cross, F. B. & Tiller, E. H. (1998), 'Judicial partisanship and obedience to legal doctrine: Whistleblowing on the federal courts of appeals', *The Yale Law Journal* **107**(7), 2155–2176.

Enea, M. (2009), 'Fitting linear models and generalized linear models with large data sets in r', *Statistical Methods for the Analysis of Large Datasets: book of short papers* pp. 411–414.

Farhang, S. & Wawro, G. (2004), 'Institutional dynamics on the u.s. court of appeals: Minority representation under panel decision making', *The Journal of Law, Economics, and Organization* **20**(2), 299–330.

Firth, D. (1993), 'Bias reduction of maximum likelihood estimates', *Biometrika* **80**, 27–38.

Fischman, J. B. (2015), 'Interpreting circuit court voting patterns: A social interactions framework', *The Journal of Law, Economics, and Organization* **31**(4), 808–842.

Halekoh, U., Højsgaard, S. & Yan, J. (2006), 'The r package geepack for generalized estimating equations', *Journal of Statistical Software* **15/2**, 1–11.

Hall, M. E. K. & Windett, J. H. (2016), 'Discouraging dissent: The chief judge's influence in state supreme courts', *American Politics Research* **44**(4), 682–709.

Haynie, S. L. (1992), 'Leadership and consensus on the u.s. supreme court', *The Journal of Politics* **54**(4), 1158–1169.

Holden, R., Keane, M. & Lilley, M. (2021), 'Peer effects on the united states supreme court', *Quantitative Economics* **12**(3), 981–1019.

Holland, P. & Leinhardt, S. (1981), 'An exponential family of probability for directed graphs', *Journal of the American Statistical Association* **76**, 33–60.

Joe, H. & Liu, Y. (1996), 'A model for a multivariate binary response with covariates based on compatible conditionally specified logistic regression', *Statistics & Probability Letters* **31**, 113–120.

Liang, K. & Zeger, S. (1986), 'Longitudinal data analysis using generalized linear models', *Biometrika* **73**, 13–22.

Pan, W., Louis, T. A. & Connett, J. E. (2000), 'A note on marginal linear regression with correlated response data', *The American Statistician* **54**, 191–195.

Pepe, M. & Anderson, G. (1994), 'A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data', *Communications in Statistics - Simulation and Computation* **23**, 939–951.

Peresie, J. L. (2005), 'Female judges matter: Gender and collegial decisionmaking in the federal appellate courts', *The Yale Law Journal* **114**(7), 1759–1790.

Preston, C. (1974), *Gibbs states on countable sets*, Cambridge University Press, Cambridge.

Qu, Y., Williams, G., Beck, G. & Goormastic, M. (1987), 'A generalized model of logistic regression for correlated data', *Communication of Statistics, A.* **16**, 3447–3476.

R Core Team (2021), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Ravikumar, P., Wainwright, M. & Lafferty, J. (2010), 'High-dimensional Ising model selection using l_1-regularized logistic regression', *The Annals of Statistics* **38**, 1287–1319.

Revesz, R. L. (1997), 'Environmental regulation, ideology, and the d. c. circuit.', *Virginia Law Review* **83**(8), 1717–1772.

Stefanski, L. & Boos, D. (2002), 'The calculus of m-estimation', *The American Statistician* **56**, 29–38.

Strauss, D. & Ikeda, M. (1990), 'Pseudolikelihood estimation for social networks', *Journal of the American Statistical Association* **85**, 204–212.

Wahlbeck, P. J., James F. Spriggs, I. & Maltzman, F. (1999), 'The politics of dissents and concurrences on the u.s. supreme court', *American Politics Quarterly* **27**(4), 488–514.

Walker, T. G., Epstein, L. & Dixon, W. J. (1988), 'On the mysterious demise of consensual norms in the united states supreme court', *The Journal of Politics* **50**(2), 361–389.

Ward, A., Corley, P. C. & Steigerwalt, A. (2023), *he Puzzle of Unanimity: Consensus on the United States Supreme Court*, Stanford University Press, Stanford.

Zhao, L. & Prentice, R. (1990), 'Correlated binary regression using a quadratic exponential model', *Biometrika* **77**, 642–648.

Zorn, C. J. W. (2001), 'Generalized estimating equation models for correlated data: A review with applications', *American Journal of Political Science* **45**(2), 470–490.

A Proof of Theorem 2.1

Proof. Following (2), the jth log-conditional distribution has the form

$$\log\left(\Pr\left(Y_j = y_j|\mathbf{Y}_{[j]} = \mathbf{y}_{[j]}\right)\right) = y_j\left[\left(\mathbf{e}_j\right)^\top \tilde{\Theta}\mathbf{y}^1_{t[j]}\right] + \log\left(1-\pi_j\right)$$

and thus the score function of the partial conditional distribution can be written as

$$\frac{\partial vec(\tilde{\Theta})}{\partial \boldsymbol{\theta}}\frac{\partial}{\partial vec(\tilde{\Theta})}\log\left(\Pr\left(Y_j = y_j|\mathbf{Y}_{[j]} = \mathbf{y}_{[j]}\right)\right) = G\left(\mathbf{y}^1_{[j]}\otimes\mathbf{e}_j\right)\left(y_j-\pi_j\right)$$

where $G = \partial vec(\tilde{\Theta})/\partial\boldsymbol{\theta}$. Specifically, $\mathbf{g}_{ij} = \partial vec(\tilde{\Theta})/\partial\theta_{ij} = \mathbf{e}_i\otimes\mathbf{e}_j + \mathbf{e}_j\otimes\mathbf{e}_i$, $\mathbf{g}_{ii} = \partial vec(\tilde{\Theta})/\partial\theta_{ii} = \mathbf{e}_i\otimes\mathbf{e}_i$, and $G = [\mathbf{g}_{11},\mathbf{g}_{12},\ldots,\mathbf{g}_{1m},\mathbf{g}_{22},\mathbf{g}_{23},\ldots,\mathbf{g}_{mm}]^\top$.

First, let V be symmetric and positive definite and $C\in\mathbb{R}^{m\times m}$ where $[C]_{ij} = E((Y_i-\pi_i)Y_j)$. When $i=j$, $[C]_{jj} = E(Y_j-\pi_j)^2 > 0$, and when $i\neq j$,

$$[C]_{ij} = E_{\mathbf{Y}_{[i]}}\left\{E_{Y_i|\mathbf{Y}_{[i]}}\left(Y_i-\pi_i\right)Y_j\right\} = 0$$

because $E_{Y_i|\mathbf{Y}_{[i]}}(Y_i) = \pi_i$. So C is a diagonal and positive definite matrix. Next, because

$$\begin{aligned}
&E\left(\left[\mathbf{Y}^1_{[j]}\otimes\mathbf{e}_j\right]\times\mathbf{e}_j^\top V(\mathbf{Y}-\boldsymbol{\pi})\right)\\
&= vec\left(E\left\{\mathbf{e}_j\mathbf{e}_j^\top V\left(\mathbf{Y}-\boldsymbol{\pi}\right)\left(\mathbf{Y}^1_{[j]}\right)^\top\right\}\right)\\
&= vec\left(\mathbf{e}_j\mathbf{e}_j^\top V(C - C\mathbf{e}_j\mathbf{e}_j^\top)\right)
\end{aligned}$$

the expectation of the estimating equations is

$$\begin{aligned}
E(\varphi^F(\boldsymbol{\theta};V)) &= GE\left(\sum_{j=1}^m(\mathbf{Y}^1_{[j]}\otimes\mathbf{e}_j)\times\mathbf{e}_j^\top V(\mathbf{Y}-\boldsymbol{\pi})\right)\\
&= G\ vec\left(\sum_{j=1}^m \mathbf{e}_j\mathbf{e}_j^\top VC - \mathbf{e}_j\mathbf{e}_j^\top VC\mathbf{e}_j\mathbf{e}_j^\top\right)\\
&= G\ vec\left(VC - \mathbb{D}(VC)\right)
\end{aligned}$$

where $\mathbb{D}(\cdot)$ returns a diagonal matrix whose diagonal elements are the diagonal elements of its argument.

When V is diagonal, $E(\varphi^F(\boldsymbol{\theta}; V)) = \mathbf{0}$ is trivial because VC is diagonal under this condition. So we only prove the other direction, i.e., $E(\varphi^F(\boldsymbol{\theta}; V)) = \mathbf{0}$ implies that V is diagonal. Because, for every $i, j = 1, \ldots, m$,

$$\begin{aligned}\eta_{ij} &= (\mathbf{e}_i \otimes \mathbf{e}_j)^\top vec\,(VC - \mathbb{D}(VC)) = \mathbf{e}_j^\top (VC - \mathbb{D}(VC))\,\mathbf{e}_i \\ &= [V]_{ji} \times [C]_{ii} - [\mathbb{D}(VC)]_{ji},\end{aligned}$$

we conclude that, if $i = j$, $\eta_{ii} = 0$ for every V, and if $i \neq j$, $\eta_{ij} = [V]_{ji}[C]_{ii}$ because $[\mathbb{D}(VC)]_{ij} = 0$. Recall that, the rows of matrix G are defined as $\mathbf{g}_{ij} = \partial vec(\tilde{\Theta})/\partial\theta_{ij} = \mathbf{e}_i \otimes \mathbf{e}_j + \mathbf{e}_j \otimes \mathbf{e}_i$, for $i \neq j$ and $\mathbf{g}_{ii} = \partial vec(\tilde{\Theta})/\partial\theta_{ii} = \mathbf{e}_i \otimes \mathbf{e}_i$, for $i = 1, \ldots, m$. Consequently, if $i = j$, $\mathbf{g}_{ii} = [VC]_{ii} - [\mathbb{D}(VC)]_{ii} = 0$ for arbitrary V, and if $j \neq i$, $[\mathbb{D}(VC)]_{ij} = 0$ and thus

$$\mathbf{g}_{ij}^\top vec(VC - \mathbb{D}(VC)) = \eta_{ij} + \eta_{ji} = [V]_{ij}([C]_{ii} + [C]_{jj}) = 0$$

if and only if $[V]_{ij} = 0$. In other words, V must be diagonal. □

B Proof of Theorem 2.2

Proof. Let $\mathbf{a}$ and $\mathbf{b}$ are two vectors in $\mathbb{R}^m$. Denote $[\mathbf{a} \circ \mathbf{b}]_j = [\mathbf{a}]_j \times [\mathbf{b}]_j$, $j = 1, \ldots, m$. Rewrite

$$\varphi^R(\boldsymbol{\theta}; V) = \sum_{j=1}^{m} \begin{bmatrix} \mathbf{x}_j \\ W_j \mathbf{Y}_{[j]}^0 \end{bmatrix} \mathbf{e}_j^\top V(\mathbf{Y} - \boldsymbol{\pi}) = \sum_{j=1}^{m} \begin{bmatrix} \mathbf{x}_j \\ W_j \mathbf{Y}_{[j]}^0 \end{bmatrix} \sum_{i=1}^{m} [V]_{ji}(Y_i - \pi_i).$$

Since $E(Y_j - \pi_j) = 0$, the upper part of the above corresponding to β has mean zero. Moreover, the expectation of the lower part corresponding to γ is

$$
\begin{aligned}
&E\left[W_j \mathbf{Y}^0_{[j]} \sum_{i=1}^{m} [V]_{ji}(Y_i - \pi_i)\right] = \sum_{j=1}^{m} W_j \sum_{i=1}^{m} [V]_{ji} E\left[\mathbf{Y}^0_{[j]}(Y_i - \pi_i)\right] \\
&= \sum_{j=1}^{m} W_j \sum_{i=1}^{m} [V]_{ji} E\left[\left(\mathbf{1}^0_{[j]} \circ \mathbf{Y}\right)(Y_i - \pi_i)\right] = \sum_{j=1}^{m} W_j \sum_{i=1}^{m} [V]_{ji} \left\{\mathbf{1}^0_{[j]} \circ ([C]_{ii}\mathbf{e}_i)\right\} \\
&= \sum_{j=1}^{m} W_j \left\{\mathbf{1}^0_{[j]} \circ C\mathbf{v}_j\right\} = \sum_{j=1}^{m} W_j C \left(\mathbf{1}^0_{[j]} \circ \mathbf{v}_j\right)
\end{aligned}
$$

where $\mathbf{1}$ is the m-dimensional 1 vector and $\mathbf{v}_j$ is the jth column of matrix V. The last equality holds because C is a diagonal matrix. When V is diagonal, $\mathbf{1}^0_{[j]} \circ \mathbf{v}_j = \mathbf{0}$ and hence the proof is complete. □