

# Using generalized estimating equations to estimate nonlinear models with spatial data

Weining Wang, Jeffrey M. Wooldridge, Mengshan Xu, Cuicui Lu & Chaowen Zheng

To cite this article: Weining Wang, Jeffrey M. Wooldridge, Mengshan Xu, Cuicui Lu & Chaowen Zheng (01 Nov 2024): Using generalized estimating equations to estimate nonlinear models with spatial data, *Econometric Reviews*, DOI: [10.1080/07474938.2024.2405487](https://doi.org/10.1080/07474938.2024.2405487)

To link to this article: <https://doi.org/10.1080/07474938.2024.2405487>



© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



View supplementary material [↗](#)



Published online: 01 Nov 2024.



Submit your article to this journal [↗](#)



Article views: 713



View related articles [↗](#)



View Crossmark data [↗](#)

# Using generalized estimating equations to estimate nonlinear models with spatial data

Weining Wang<sup>a</sup>, Jeffrey M. Wooldridge<sup>b</sup>, Mengshan Xu<sup>c</sup>, Cuicui Lu<sup>d</sup>, and Chaowen Zheng<sup>e</sup>

<sup>a</sup>School of Economics, University of Bristol, Bristol, UK; <sup>b</sup>Department of Economics, Michigan State University, East Lansing, Michigan, USA; <sup>c</sup>Department of Economics, University of Mannheim, Mannheim, Germany; <sup>d</sup>School of Innovation and Entrepreneurship, Shandong University, Qingdao, Shandong Province, China; <sup>e</sup>Department of Economics, University of Southampton, Southampton, UK

## ABSTRACT

We study the estimation of nonlinear models with cross-sectional data using two-step generalized estimating equations within the quasi-maximum likelihood estimation framework. To improve efficiency, we propose a grouped estimator that accounts for potential spatial correlation in the underlying innovations of nonlinear models. Under mild weak dependence assumptions, we provide results on estimation consistency and asymptotic normality. Monte Carlo simulations demonstrate the efficiency gain of our approach compared to various estimation methods. Finally, we apply the proposed approach to examine the role of cultural distance in an extended gravity equation using international trade data from China. Compared to existing methods, our approach yields estimates with smaller standard errors and reinforces the hypothesis that both cultural and geographical distances significantly negatively influence international trade.

## ARTICLE HISTORY

15 December 2023  
10 September 2024

## KEYWORDS

Efficiency gain; nonlinear models; quasi-maximum likelihood estimation; spatial dependence; the gravity equation

## JEL CODES:


C13, C21, C35, C51

## 1. Introduction

In empirical economics and social studies, there are many examples of discrete (noncontinuous) data that exhibit spatial or cross-sectional correlations due to the "distance" in a space. This "distance" can be geographical, economic, cultural, institutional, and so on. Arbia and Billé (2018) recently summarized the existing literature on spatial discrete choice models. These nonlinear spatial models are used to study the effect of nearby individuals due to various effects, such as spillover effect, neighborhood effect, or peer effect. For example, the number of patents a firm receives can be affected by other nearby firms (Bloom, Schankerman, and Van Reenen, 2013); an individual's decision on whether to own stocks is affected by the average stock market participation of the individual's community (e.g., Brown et al., 2008); a student's academic performance is affected by his or her roommate (e.g., Sacerdote, 2001; and Angrist, 2014).

Nonlinear models are more appropriate than linear models for discrete (noncontinuous) response data because they better handle the data's bounded nature and allow the partial effect of any explanatory variable to vary (see, e.g., Chapter 17 in Wooldridge (2020)). There are many studies on the theoretical properties of nonlinear spatial autoregressive models (see, e.g., Xu and Lee, 2015a, 2015b), wherein the dependent variable appears on the right-hand side of the equation. These models are also popular in empirical research, as they can be used to model relationships between players in a game (de Paula, 2013)

**CONTACT** Mengshan Xu  [mengshan.xu@uni-mannheim.de](mailto:mengshan.xu@uni-mannheim.de)  Department of Economics, University of Mannheim, L7, 3-5, 68161 Mannheim, Germany; Cuicui Lu  [lucucui@sdu.edu.cn](mailto:lucucui@sdu.edu.cn)  School of Innovation and Entrepreneurship, Shandong University, Qingdao, Shandong Province, 266200 China.

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/07474938.2024.2405487>.

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

and social interactions in a network context (Lee, Liu, and Lin, 2010). Moreover, in a model with spatially correlated dependent variables, the spatial autocorrelation parameter allows for assessing the direction and strength of the effect (e.g., Gagliardini, Ossola, and Scaillet, 2020; Elhorst, 2014). A comprehensive summary and more examples of spatial autoregressive (SAR) models can be found in Arbia and Billé (2018), Arbia (2016), and Baltagi, Egger, and Kesina (2016).

In contrast, our research focuses on the estimation of nonlinear models with spatial errors, where spatial dependence is modeled between error terms (and explanatory variables) for different individuals. On the one hand, researchers have a keen interest in understanding various econometric issues for models with solely spatial errors. For example, Baltagi, Song, and Koh (2003) studied statistical testing problems in panel models with only spatial error correlation, and Kapoor, Kelejian, and Prucha (2007) investigate consistent and efficient estimation for panels with only spatial error correlations. On the other hand, there are also many empirical examples where models with spatial errors may be more appropriate than spatial models like SAR. For instance, when the dependent variables are country-wise GDPs, it may not be intuitive to draw causal links directly among GDPs. Instead, it is more meaningful to understand the spillover effect from other variables or unobserved characteristics/shocks. Another example is when the dependent variable is the number of murders; it is difficult to explain the direct interaction between murder counts across different counties. Instead, the underlying (unobserved) correlated social and economic characteristics are more likely to determine these numbers, and it makes more sense to study the spillover effect via a spatial error model. Spatial error models are also useful when investigating spatial heterogeneity. When only a single observation for each region is available, it is impossible to estimate the unobserved individual heterogeneous effects directly. One way to tackle this issue, as is commonly done in the spatial literature, is to assume these effects are similar to those of neighboring units and model them via spatially correlated errors (LeSage and Pace, 2009). Thus, we shall focus exclusively on the nonlinear models without spatial lags. More examples of spatially correlated error models can be found in Graham (2008) and Carrell, Sacerdote, and West (2013).

There has been a sizable literature on models accounting for nonlinear spatially dependent errors. Although the dependence structure of the underlying error is generally unknown in a spatial dataset, many methods do not allow misspecifications that could lead to misleading results. For example, if the joint distribution of the variables is misspecified, the maximum likelihood estimator (MLE) is not consistent in general. One of the alternative methods is partial-maximum likelihood estimation (PMLE), which only uses marginal distributions. Wang, Iglesias, and Wooldridge (2013) use a bivariate Probit PMLE to improve the estimation efficiency of a spatial error model. Their approach requires correctly specifying the marginal distribution of the binary response variable conditional on the covariates and distance measures (a distance measure is how one defines the distances between observations). Another method is quasi-maximum likelihood estimation (QMLE). Using a density that belongs to a linear exponential family (LEF), QMLE is consistent if we correctly specify the conditional mean, while other features of the density can remain misspecified (Gourieroux, Monfort, and Trognon, 1984). Lee (2004) derives asymptotic distributions of QMLE for SAR models without assuming normal distributions.

Different from QMLE and PMLE, we shall adopt a method that produces a consistent estimator and allows moderate misspecification of the underlying dependence structure. Moreover, it has favorable efficiency performance when we are exploiting some given information on the dependence structure. In this article, we adopt a generalized estimating equations (GEE) method for spatial data sets. The GEE approach is one of the QMLE methods since it takes a specific form of the maximum likelihood score equation for a multivariate Gaussian distribution. It can be used to account for serial correlation and thus can achieve more efficient estimators, and it is adopted to estimate the parameters of a generalized linear model with a possible unknown correlation between outcomes (Liang and Zeger, 1986). The spatial GEE method proposed in our paper further relaxes distributional assumptions in the literature. We assume that the mean function is correctly specified, and we choose a working variance-covariance matrix, which is defined as a prespecified variance-covariance matrix. It may not be the same as the true one since the true variance-covariance matrix is generally unknown. Under mild regularity conditions, parameter estimates from the GEE are consistent even if this working variance-covariance structure is misspecified.

Though the specification of the working variance-covariance matrix does not affect the consistency, it does play a significant role in the efficiency of the estimator. We propose to work with data with nature groups or some pre-given grouping information: Presumably, within the group there exists stronger dependence, and between-group individuals are less related to each other. For example, in our cultural distance applications, we group the data according to the natural geographical information. While this group partition may make our model very similar to panel data, it is only imposed on the error term to achieve estimation efficiency. Therefore, our model is of a very different nature from the panel data model. As one may want to use techniques developed in panel data such as the random effects method to achieve estimation efficiency, they are not directly applicable to our model as there are only cross-section individuals in our model. How to model and estimate the cross-sectional dependence is much more involved than for temporal dependence. As one of the main contributions of this paper, we adapt the theories developed in Jenish and Prucha (2012, 2007) for our analysis. Furthermore, we also examine the conditions under which efficiency gains can be achieved. Following the intuition of a nonlinear weighted least square estimator, in Section 2.4, we show by some nontrivial algebra that the extent of efficiency improvement depends on how closely the working variance-covariance matrix approximates the true one. This means that the adopted group structure should appropriately reflect the underlying dependence.

There are several studies applying grouping or blocking structures for estimating nonlinear GEE models. Rao Chaganty and Joe (2004), Lin and Clayton (2005), and Oman et al. (2007) use estimating equations for the binary response model, assuming that blocks are independently chosen. Adegboye, Leung, and Wang (2018) analyze spatial data by considering different correlation structures. Our setup is different from the research mentioned above: While their estimating equations focus on the correlations of all individuals, our article assumes near-epoch dependence (NED) among all observations but applies group partition when estimating the parameters. The proposed method only picks up within-group correlations while, admittedly, there are correlations between individuals belonging to different groups. We use the QMLE that ignores correlations within groups as the initial estimator for a two-step GEE and show the efficiency gain of this GEE estimator.

Our method extends the QMLE proposed in Lee (2004) to nonlinear cases, who discusses a SAR model that includes a spatially lagged dependent variable as an additional right-hand side regressor. Our technique partly overlaps with the method in Xu and Lee (2015a). While they investigate a Tobit model with a spatially lagged dependent variable as an additional regressor, we cover a more general class of models, including those for count data with Poisson regression. Moreover, we theoretically show the consistency of our GEE approach within the QMLE framework in a spatial data setting. To derive the asymptotics for the GEE estimator, we use a uniform law of large numbers (ULLN) and a central limit theorem (CLT). Our theoretical development is different from Conley (1999) and Jenish and Prucha (2012, 2007), who derive the asymptotics for GMM estimators of spatial processes and for mixing or NED spatial processes, respectively. While their hyperassumptions are directly imposed on the outcome variables, ours are imposed on the latent innovations. Moreover, it is important to acknowledge that GEE can be perceived as a specific instance of Z-estimation, as highlighted in Chapter 5 of Van der Vaart (2000). Our investigation has been conducted to understand how the near-epoch dependence property of the underlying processes contributes to the proof of the estimator's asymptotic properties. Finally, we provide a consistency proof of a proposed semiparametric estimator for the variance-covariance matrix.

To summarize, we contribute to the literature in four aspects. First, we develop a simple GEE method for a general class of nonlinear models, which uses less distributional assumptions by only specifying the conditional mean for spatially dependent data. The proposed technique simply groups data and imposes weights to adjust for the group-wise dependence, and we model the spatial correlation in the underlying innovations rather than in the dependent variable. Second, we focus on the aspects of the efficiency gain with grouped data, in addition to the consistency of the estimation. Further, we provide a condition on the working variance-covariance matrix that ensures efficiency gains over QMLE. Third, we prove the asymptotic properties of our method by applying a ULLN and a CLT to the spatial GEE estimator. Finally, we show in the simulation study that the proposed GEE method using spatial correlation has

considerable efficiency gain for two types of data: count and binary response, and it is robust to moderate group misspecification.

The article is organized as follows. In [Section 2](#), the GEE methodology under the spatial data context is proposed. [Section 3](#) looks in detail at a Poisson model and a negative Binomial II model for count data with a multiplicative spatial error term, and we further study a Probit model for binary response data with spatial correlation in the latent error term. In [Section 4](#), a series of assumptions are given, under which we establish theorems on consistency and asymptotic normality of the spatial GEE estimator and provide a consistent estimator for its variance-covariance matrix. [Section 5](#) contains Monte Carlo simulation results that compare the efficiency of different estimation methods for the nonlinear models explored in the previous sections. [Section 6](#) includes an application to study an extended gravity equation on trade volume between China and its trade partners using country-product level data. [Section 7](#) concludes the article. The proofs and other technical details are provided in supplementary materials.

## 2. Methodology

### 2.1. Notation and definition

We first lay out the basic notations for our methodology. We delay important assumptions and asymptotic results until [Section 4](#).

Let  $\theta \in \Theta \subset \mathbb{R}^p$  be the parameter of interest in the conditional mean, and  $\gamma \in \Gamma \subset \mathbb{R}^q$  be the nuisance variance parameter involved in the conditional variance.  $\Theta \times \Gamma$  is a compact set, and  $(\theta^0, \gamma^0)$  are the true parameter values. We shall note that  $\theta$  involved in the conditional variance will also be treated as a nuisance parameter. Let the group index be  $g \in \{1, \dots, G\}$ , and  $L_g$  be the number of observations in group  $g$ .  $L_g$  can be different and bounded by a constant. For simplicity, we assume  $L_g = L$  for all  $g$ . Notice that the group structure is only for error terms and the parameters of interest are homogeneous across groups (see [Graham, 2008](#) for a similar setting). Otherwise, the model could be estimated by group-wise analysis. Let  $d_{ij}$  denote the distance between observations  $i$  and  $j$ , and let  $d_{gh}$  denote the distance between group  $g$  and  $h$ .

In this article, we consider spatial processes located on an unevenly spaced lattice  $D \subseteq \mathbb{R}^d$  for  $d \geq 1$ . Moreover,  $D_G$  denotes the lattice containing group locations, and each group location is denoted as vectorizing the elements in  $B_g$ , where  $B_g$  is the associated set of locations within the group  $g$ .  $D_n$  is defined similarly for each observation. Let the total number of groups be  $|D_G| = G$ , and the total number of observations is  $|D_n| = n$ . Let  $|U|$  denote the cardinality of a finite subset  $U \subseteq D$ .

### 2.2. The generalized estimating equations methodology

We extend the GEE methodology in [Liang and Zeger \(1986\)](#) to the estimation of nonlinear spatial data, using a two-step procedure that first estimates the working correlation matrix and then solves the generalized estimating equations. In our approach, we divide spatially correlated cross-sectional data into groups, allowing arbitrary strong dependence within each group but requiring the between-group dependence to diminish (in the sense of  $\alpha$ -mixing that will be defined in [Section 4.2](#)). We assume that the division of groups is exogenous, that is, how groups are divided does not affect outcomes once controlled for explanatory variables. The division of groups can be based on, for example, geographical properties or researcher-defined economic and social relationships. There are two extreme cases of group size. The first case is when the group size is one, where the resulting estimators ignore all of the pairwise correlations. The second case is when the group size is  $n$ , which means we are using all of the pairwise information. If the group size is not equal to one or  $n$ , we have a "partial" estimator. By "partial," we mean that instead of using full information, we only use the information within the same groups. Note that in our settings, the number of groups  $G \rightarrow \infty$ , and the group size  $L$  are assumed to be fixed. Similar settings are also maintained in many existing studies; see e.g., [Section 20.3.1](#) in [Wooldridge \(2010\)](#) and [Wooldridge \(2003\)](#). Moreover, these settings are compatible with many applications. For example, the

group can represent classrooms, schools, families, or firms for which the number could be very large, and the group members are generally fixed. In an empirical study investigating the impact of class size on student achievement by Carter, Schnepel, and Steigerwald (2017) (see also Krueger, 1999), there are 318 groups (classrooms), and the number of students in each classroom is no more than 27.

Consider the following nonlinear model for the  $i$ -th observation  $y_i$ :

$$y_i = m(\mathbf{x}_i; \theta^0) + u_i, \quad (1)$$

where the dependent variable  $y_i$  can be continuous or discrete,  $m(\mathbf{x}_i; \theta^0)$  is the conditional expectation function,  $\mathbf{x}_i$  is a  $1 \times p$  row vector of independent variables which can be continuous, discrete, or a combination,  $\theta_0$  is the parameter of interest, and  $u_i$  is the unobserved error term.

The model for  $g$ -th group observations  $\mathbf{y}_g$  then could be written as

$$\mathbf{y}_g = \mathbf{m}(\mathbf{x}_g; \theta^0) + \mathbf{u}_g. \quad (2)$$

where  $\mathbf{y}_g$  and  $\mathbf{u}_g$  are both  $L \times 1$  vectors, including all dependent variables and error terms in group  $g$ ,  $\mathbf{m}(\mathbf{x}_g; \theta^0)$  (abbreviated as  $\mathbf{m}_g(\theta^0)$ ) is an  $L \times 1$  vector of conditional mean functions, and  $\mathbf{x}_g$  is an  $L \times p$  matrix, including all independent variables in  $g$ -th group.

Note that for each individual in group  $g$ ,  $\mathbf{x}_i$  is allowed to contain explanatory variables from other individuals within the same group. To explain this with an example, we can write an explicit expectation of observation  $i$  as

$$E(y_i | \mathbf{x}_i) = m(\mathbf{x}_i; \theta^0), \quad (3)$$

where  $\mathbf{x}_i = [\mathbf{z}_i, \mathbf{z}_{(-i)}]$  for  $i = 1, 2, \dots, n$ , and  $\mathbf{z}_{(-i)}$  is some weighted value of other group members' explanatory variables, capturing the exogenous interaction/spillover effects among the independent variables with different intensities/weights. We denote the conditional variance-covariance matrix of  $\mathbf{y}_g$  as  $\mathbf{W}_g \stackrel{\text{def}}{=} \text{Cov}(\mathbf{y}_g, \mathbf{y}_g | \mathbf{x}_g) = E(\mathbf{y}_g \mathbf{y}_g^\top | \mathbf{x}_g) - E(\mathbf{y}_g | \mathbf{x}_g) E(\mathbf{y}_g | \mathbf{x}_g)^\top$ , which is unknown in most cases. Usually, for the conditional mean function,  $\theta^0 \in \Theta \subset \mathbb{R}^p$  is the main parameter of interest. We can parameterize the corresponding weight matrix  $\mathbf{W}_g$  by  $\mathbf{W}_g(\theta, \gamma)$  with  $\gamma \in \Gamma \subset \mathbb{R}^q$  and  $\theta$  as (first-stage) nuisance parameters, which are involved only in the estimation of the variance-covariance matrix.

The objective function for group  $g$  and the whole sample are given as follows:

$$q_g(\theta, \gamma) = (\mathbf{y}_g - \mathbf{m}_g(\theta))^\top \mathbf{W}_g^{-1}(\theta, \gamma) (\mathbf{y}_g - \mathbf{m}_g(\theta)), \quad (4)$$

$$Q_G(\theta, \gamma) = G^{-1} \sum_{g=1}^G q_g(\theta, \gamma). \quad (5)$$

Note that the objective function of GEE only uses group-wise information. Considering that  $\gamma$  is a nuisance parameter, the quasi-score equation for estimating  $\theta$  is then defined as follows:

$$\mathbf{S}_G(\theta, \gamma) = \frac{1}{G} \sum_g \nabla \mathbf{m}_g(\theta)^\top \mathbf{W}_g^{-1}(\theta, \gamma) [\mathbf{y}_g - \mathbf{m}_g(\theta)], \quad (6)$$

where  $\nabla \mathbf{m}_g(\theta)$  is the gradient of  $\mathbf{m}_g(\theta)$  with respect to  $\theta$ . The above quasi-score equation is similar to Eq. (6) in Liang and Zeger (1986). In this regard, the score function  $\mathbf{S}_G(\theta, \gamma)$  is parameterized by  $\{\theta, \gamma\}$ , and the choice of  $\theta$  and  $\gamma$  involved in  $\mathbf{W}_g^{-1}(\theta, \gamma)$  would affect the estimation efficiency of  $\theta^0$ . In practice, we obtain the first-step estimator  $\{\check{\theta}, \check{\gamma}\}$ , and then estimate  $\theta^0$  with  $\mathbf{W}_g^{-1}(\check{\theta}, \check{\gamma})$ . Namely, we consider the GEE estimator  $\hat{\theta}$  that solves the following equation:

$$\frac{1}{G} \sum_{g=1}^G \nabla \mathbf{m}_g(\hat{\theta})^\top \mathbf{W}_g^{-1}(\check{\theta}, \check{\gamma}) [\mathbf{y}_g - \mathbf{m}_g(\hat{\theta})] = 0. \quad (7)$$

This equation is similar to the Eq. (7) in Liang and Zeger (1986), who solve  $\theta$  given a pre-estimator of the nuisance parameter. It is worth noting that we need some identification assumptions (i.e., Assumption A.5 in Section 4.3) to ensure the existence and uniqueness of the solution.

We denote the population version of loss as  $Q_\infty(\theta, \gamma) \stackrel{\text{def}}{=} \lim_{G \rightarrow \infty} G^{-1} \sum_g \text{E} q_g(\theta, \gamma)$  and  $\mathbf{S}_\infty(\theta, \gamma) \stackrel{\text{def}}{=} \lim_{G \rightarrow \infty} \text{E} \mathbf{S}_G(\theta, \gamma)$ . To make a difference between the first step plug-in nuisance parameter and the parameter of interest, we rewrite the score into

$$\mathbf{S}_G(\check{\theta}, \check{\gamma}, \theta) = \frac{1}{G} \sum_g \nabla \mathbf{m}_g(\theta)^\top \mathbf{W}_g^{-1}(\check{\theta}, \check{\gamma}) [\mathbf{y}_g - \mathbf{m}_g(\theta)].$$

Correspondingly,  $Q_G(\check{\theta}, \check{\gamma}, \theta) = G^{-1} \sum_g (\mathbf{y}_g - \mathbf{m}_g(\theta))^\top \mathbf{W}_g^{-1}(\check{\theta}, \check{\gamma}) (\mathbf{y}_g - \mathbf{m}_g(\theta))$ . Frequently, we restrict our attention to the exponential family, which embraces many distributions, such as the Bernoulli, Poisson, and Gaussian distributions. Now, we link this estimation method with a QMLE framework. We suppress the parameter  $\gamma$  for a moment. We assume that the probability density function  $f(\mathbf{y}_g | \mathbf{x}_g; \theta)$  is in the LEF (See details of the exponential family in Section S3 of supplementary materials). For instance, when there is only one observation for each group, i.e., we do not account for the spatial covariance, a characterization of QMLE in LEF is by the following individual score function:

$$\mathbf{s}_i(\theta) = \nabla m(\mathbf{x}_i; \theta)^\top \{y_i - m(\mathbf{x}_i; \theta)\} / v(m(\mathbf{x}_i; \theta)), \quad (8)$$

where  $\nabla m(\mathbf{x}_i; \theta)$  is the  $1 \times p$  gradient of the mean function and  $v(m(\mathbf{x}_i; \theta))$  is the conditional variance function associated with the chosen LEF density. For Bernoulli distribution, we have  $v(m(\mathbf{x}_i; \theta)) = m(\mathbf{x}_i; \theta)(1 - m(\mathbf{x}_i; \theta))$ , and for Poisson distribution,  $v(m(\mathbf{x}_i; \theta)) = m(\mathbf{x}_i; \theta)$ . Note that consistent estimation of parameter  $\theta^0$  could be obtained based on (8). While it accounts for potential heteroscedasticity, this estimator might not be the most efficient since it overlooks potential cross-sectional correlations between observations.

Comparably, recall the score function for GEE can be written as

$$\mathbf{S}_G(\theta, \gamma) = \frac{1}{G} \sum_g s_g(\theta, \gamma) = \frac{1}{G} \sum_g \nabla \mathbf{m}_g^\top(\theta) \mathbf{W}_g^{-1}(\theta, \gamma) [\mathbf{y}_g - \mathbf{m}_g(\theta)],$$

where  $s_g(\theta, \gamma) = \nabla \mathbf{m}_g^\top(\theta) \mathbf{W}_g^{-1}(\theta, \gamma) [\mathbf{y}_g - \mathbf{m}_g(\theta)]$  is the  $p \times 1$  vector of the score for the group  $g$  with  $\mathbf{W}_g^{-1}(\theta, \gamma)$  accounting for group dependence. Accordingly, we denote the Hessian matrix as  $\mathbf{H}_G(\theta, \gamma) = \nabla_\theta \mathbf{S}_G(\theta, \gamma)$ , and  $h_g(\theta, \gamma) = \nabla_\theta s_g(\theta, \gamma)$  as the  $p \times p$  matrix of Hessian for the group  $g$ . We also define  $\mathbf{H}_\infty(\theta, \gamma) = \lim_{G \rightarrow \infty} \text{E}[\mathbf{H}_G(\theta, \gamma, \theta)]$ .

### 2.3. The first-step estimation of the weight matrix

In this subsection, we demonstrate one way to find an estimator for  $\gamma$  in  $\mathbf{W}_g(\theta, \gamma)$ , which can be written as

$$\mathbf{W}_g(\theta, \gamma) = \mathbf{V}(\mathbf{x}_g; \theta)^{1/2} \Omega_g(\gamma) \mathbf{V}(\mathbf{x}_g; \theta)^{1/2}, \quad (9)$$

where  $\Omega_g(\gamma)$  is the  $L \times L$  correlation matrix for the group  $g$ , and  $\mathbf{V}(\mathbf{x}_g; \theta)$  is the  $L \times L$  diagonal matrix that only contains variances of  $\mathbf{y}_g - \mathbf{m}_g(\mathbf{x}_g, \theta^0)$ :

$$\mathbf{V}(\mathbf{x}_g; \theta) = \begin{pmatrix} v_{g1} & 0 & \cdots & 0 \\ 0 & v_{g2} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & v_{gL} \end{pmatrix}, \quad (10)$$

where  $\{g_1, g_2, \dots, g_L\}$  is a subset of  $\{i\}_{i=1}^n$ , indicating the members of the group  $g$ . The  $l$ -th element on the diagonal is the variance  $v_{gl} := \text{Var}(\mathbf{y}_{gl} | \mathbf{x}_{gl})$  for  $l$ -th individual in the group  $g$ .  $\mathbf{y}_{gl}$  is the  $l$ -th element in

the vector  $\mathbf{y}_g$ , and  $\mathbf{x}_{gl}$  is the  $l$ -th row in  $\mathbf{x}_g$ . Moreover, the correlation matrix is

$$\Omega_g(\gamma) = \begin{pmatrix} 1 & \pi_{g12} & \cdots & \pi_{g1L} \\ \pi_{g21} & 1 & & \vdots \\ \vdots & & \ddots & \pi_{g(L-1)L} \\ \pi_{gL1} & \cdots & \pi_{gL(L-1)} & 1 \end{pmatrix}. \quad (11)$$

Let  $d_{glm}$  be the distance between the  $l$ -th and the  $m$ -th observations in group  $g$ . An example of a parametrization of the  $(l, m)$ -th element (with  $l \neq m$ ) of  $\Omega_g(\gamma)$  would be  $\pi_{glm} = \rho \left(1 - \check{d}_{glm}\right)$  (see e.g. Cressie, 2015) with  $\gamma \equiv \rho$  being the spatial correlation parameter subject to  $0 \leq \rho \leq 1$ , and  $0 \leq \check{d}_{glm} \leq 1$  is the normalized  $d_{glm}$ . This choice is feasible as it assumes a linear decaying rate of spatial dependence.

Given a parametrization, we discuss the way to estimate  $\gamma$ . Let  $\check{\theta}$  be the first-step QMLE estimator,  $\check{u}_i = y_i - m(\mathbf{x}_i; \check{\theta})$  be the first-step residual, and  $\check{v}_i = v(m(\mathbf{x}_i; \check{\theta}))$  be the fitted variance of the individual  $i$  corresponding to the chosen LEF density. We also denote  $\check{r}_i = \check{u}_i / \sqrt{\check{v}_i}$  as the standardized residual. Recall that  $\{g_1, g_2, \dots, g_L\}$  is a subset of  $\{i\}_{i=1}^n$ , indicating the members in the group  $g$ . Let  $\check{\mathbf{r}}_g = (\check{r}_{g_1}, \check{r}_{g_2}, \dots, \check{r}_{g_L})^\top$ . Then  $\check{\mathbf{r}}_g \check{\mathbf{r}}_g^\top$  is the estimated sample correlation matrix for the group  $g$ . Let  $\mathbf{e}_g(\check{\theta})$  be a vector containing  $L(L-1)/2$  different elements of the lower (or upper) triangle of  $\check{\mathbf{r}}_g \check{\mathbf{r}}_g^\top$ , excluding the diagonal elements. Let  $\mathbf{z}_g(\gamma)$  be the vector containing the elements in  $\Omega_g(\gamma)$  corresponding to the same entries of elements in  $\check{\mathbf{r}}_g \check{\mathbf{r}}_g^\top$ . We can follow Prentice (1988), and find a consistent estimator for  $\gamma$  by solving

$$\check{\gamma} = \mathbf{argmin}_{\gamma \in \Gamma} \sum_g (\mathbf{e}_g(\check{\theta}) - \mathbf{z}_g(\gamma))^\top (\mathbf{e}_g(\check{\theta}) - \mathbf{z}_g(\gamma)). \quad (12)$$

**Remark 1.** As will be shown in [Theorem 1](#) in [Section 4](#), our proposed GEE estimator of  $\theta^0$  is always consistent as long as the conditional mean is correctly specified, even when the underlying dependence structure is misspecified or  $\gamma$  is not consistently estimated. The misspecification of the dependence structure and the estimation of  $\gamma$  only affect the efficiency but not the consistency. In the following [Section 2.4](#), we adopt an honest approach and provide conditions under which our estimator could achieve efficiency gains over a QMLE estimator that does not consider any dependence, suggesting that our proposed GEE method can accommodate a moderate misspecification of underlying dependence structure. Inspired by the referees, to understand how close the specification is from the true (unknown) structure, we may also employ some nonparametric/resampling methods (e.g., bootstrap) to recover the true structure. The theoretical properties of these estimators are, however, unclear and may rely on some stringent conditions, and we leave it as an interesting future research topic.

## 2.4. Conditions for the improvement

The efficiency gain comes from the fact that the proposed GEE method accounts for the spatial dependence that QMLE or PMLE methods fail to (or are not flexible enough to) consider. The limitation of QMLE can be due to the non linearity of the model (the example in [Section 3.1](#)) or a latent spatial error term (the example in [Section 3.2](#)).

In this section, we provide a condition on the relationship between the working variance-covariance matrix and the true variance-covariance matrix, and this condition ensures efficiency improvement in comparison to a QMLE estimator. Namely, if the working variance-covariance matrix is sufficiently close to the true one, we can achieve an efficiency gain.

Recall that the parameter dimension of  $\theta$  is  $p$ . Denote  $E(u_i^2 | \mathbf{x}_{1:n}) = v_i^2$  where  $\mathbf{x}_{1:n}$  is an  $n \times p$  matrix of all the independent variables, and  $\nabla \mathbf{m}^\top$  (a  $n \times p$  matrix) denotes a stack form of  $\nabla \mathbf{m}_g(\theta^0)$ . For simplicity, we suppress  $\theta^0$  and write  $\nabla \mathbf{m}_g(\theta^0)$  as  $\nabla \mathbf{m}_g$  from now on. We also let  $\mathbf{W}$  denote a block



diagonal matrix, where its  $g$ -th block element is  $\mathbf{W}_g$  (an abbreviation of matrix  $\mathbf{W}_g(\theta^0, \gamma^0)$  defined in the previous section).  $\Omega^{-1}$  is a diagonal matrix, where its  $i$ -th element is  $v_i^{-2}$ . We compare the QMLE estimator  $\hat{\theta}_{ug}$  with the proposed grouped estimator  $\hat{\theta}_{GEE}$ . Denote the gradient of  $m(\mathbf{x}_i; \theta^0)$  as  $\nabla \mathbf{m}(\mathbf{x}_i; \theta^0)$ . The matrices involved in the variance of  $\hat{\theta}_{ug}$  would be

$$\Sigma_A := \sum_{i=1}^n (\nabla \mathbf{m}(\mathbf{x}_i; \theta^0))^\top v_i^{-2} \nabla \mathbf{m}(\mathbf{x}_i; \theta^0) = \nabla \mathbf{m}^\top \Omega^{-1} \nabla \mathbf{m},$$

and

$$\Sigma_B := \sum_{i=1}^n \sum_{j=1}^n \nabla \nabla \mathbf{m}(\mathbf{x}_i; \theta^0)^\top v_i^{-2} \mathbb{E}(u_i u_j | \mathbf{x}_{1:n}) v_j^{-2} \nabla \mathbf{m}(\mathbf{x}_j; \theta^0) = \nabla \mathbf{m}^\top \Omega^{-1} V \Omega^{-1} \nabla \mathbf{m},$$

where  $V$  denotes the true variance-covariance matrix, and its  $ij$ -th element is  $\mathbb{E}(u_i u_j | \mathbf{x}_{1:n})$ .

Thus, the asymptotic variance-covariance matrix of  $\hat{\theta}_{ug}$  is of the following form:

$$\text{Avar}(\hat{\theta}_{ug} | \mathbf{x}_{1:n}) = \Sigma_A^{-1} \Sigma_B \Sigma_A^{-1}.$$

As a comparison, the asymptotic variance of the grouped estimator (i.e.,  $\hat{\theta}_{GEE}$ ) is characterized by the following two matrices:

$$\Sigma_{G,A} := \sum_g \nabla \mathbf{m}_g^\top \mathbf{W}_g^{-1} \nabla \mathbf{m}_g = \nabla \mathbf{m}^\top \mathbf{W}^{-1} \nabla \mathbf{m},$$

$$\Sigma_{G,B} := \sum_g (\nabla \mathbf{m}_g^\top \mathbf{W}_g^{-1} \mathbb{E}(\mathbf{u}_g \mathbf{u}_g^\top | \mathbf{x}_{1:G}) \mathbf{W}_g^{-1} \nabla \mathbf{m}_g) = \nabla \mathbf{m}^\top \mathbf{W}^{-1} V \mathbf{W}^{-1} \nabla \mathbf{m}.$$

where recall  $\mathbf{u}_g = \mathbf{y}_g - \mathbf{m}_g(\theta^0)$ .

The conditional asymptotic variance of  $\hat{\theta}_{GEE}$  is denoted as

$$\text{Avar}(\hat{\theta} | \mathbf{x}_{1:G}) = \Sigma_{G,A}^{-1} \Sigma_{G,B} \Sigma_{G,A}^{-1}.$$

A desirable condition is to ensure the positive definiteness of the matrix,

$$\Sigma_A^{-1} \Sigma_B \Sigma_A^{-1} - \Sigma_{G,A}^{-1} \Sigma_{G,B} \Sigma_{G,A}^{-1}.$$

It shall be noted that if  $\mathbf{W}_g$  is correctly specified, we have  $\mathbb{E}(\mathbf{u}_g \mathbf{u}_g^\top | \mathbf{x}_g) = \mathbf{W}_g$ . Thus, we have  $\Sigma_{G,B} = \sum_g \nabla \mathbf{m}_g^\top \mathbf{W}_g^{-1} \nabla \mathbf{m}_g = \Sigma_{G,A}$ . It follows that the above estimator attains the lower bound, i.e.,

$$\Sigma_{G,A}^{-1} \Sigma_{G,B} \Sigma_{G,A}^{-1} = \Sigma_{G,A}^{-1}.$$

We assume that the true matrix can be expressed as  $V = \mathbf{W} + \delta B$ , where  $\delta$  is some small enough positive constant and  $B$  is a symmetric matrix. This basically assumes that  $\mathbf{W}$  is equal to the true matrix  $V$  plus a perturbation. This corresponds to the case of moderate misspecification.

To understand the improvement in the variance of the GEE estimator, we express its variance as

$$\Sigma_{G,A}^{-1} \Sigma_{G,B} \Sigma_{G,A}^{-1} = (\nabla \mathbf{m}^\top \mathbf{W}^{-1} \nabla \mathbf{m})^{-1} + \delta (\nabla \mathbf{m}^\top \mathbf{W}^{-1} \nabla \mathbf{m})^{-1} (\nabla \mathbf{m}^\top \mathbf{W}^{-1} B \mathbf{W}^{-1} \nabla \mathbf{m}) (\nabla \mathbf{m}^\top \mathbf{W}^{-1} \nabla \mathbf{m})^{-1}.$$

We shall see that the variance of the ungrouped QMLE estimator is

$$\begin{aligned} \Sigma_A^{-1} \Sigma_B \Sigma_A^{-1} &= (\nabla \mathbf{m}^\top \Omega^{-1} \nabla \mathbf{m})^{-1} (\nabla \mathbf{m}^\top \Omega^{-1} \mathbf{W} \Omega^{-1} \nabla \mathbf{m}) (\nabla \mathbf{m}^\top \Omega^{-1} \nabla \mathbf{m})^{-1} \\ &\quad + \delta (\nabla \mathbf{m}^\top \Omega^{-1} \nabla \mathbf{m})^{-1} (\nabla \mathbf{m}^\top \Omega^{-1} B \Omega^{-1} \nabla \mathbf{m}) (\nabla \mathbf{m}^\top \Omega^{-1} \nabla \mathbf{m})^{-1}. \end{aligned}$$

Now, we aim to find a condition such that  $\Sigma_A^{-1} \Sigma_B \Sigma_A^{-1} - \Sigma_{G,A}^{-1} \Sigma_{G,B} \Sigma_{G,A}^{-1}$  is positive definite.

First, note that

$$\begin{aligned} & (\nabla \mathbf{m}^\top \Omega^{-1} \nabla \mathbf{m})^{-1} (\nabla \mathbf{m}^\top \Omega^{-1} \mathbf{W} \Omega^{-1} \nabla \mathbf{m}) (\nabla \mathbf{m}^\top \Omega^{-1} \nabla \mathbf{m})^{-1} - (\nabla \mathbf{m}^\top \mathbf{W}^{-1} \nabla \mathbf{m})^{-1} \\ &= (\nabla \mathbf{m}^\top \Omega^{-1} \nabla \mathbf{m})^{-1} (\nabla \mathbf{m}^\top \Omega^{-1} \mathbf{W}^{1/2}) (I - \mathbf{W}^{-1/2} \nabla \mathbf{m} (\nabla \mathbf{m}^\top \mathbf{W}^{-1} \nabla \mathbf{m})^{-1} \\ & \nabla \mathbf{m}^\top \mathbf{W}^{-1/2}) (\mathbf{W}^{1/2} \Omega^{-1} \nabla \mathbf{m}) (\nabla \mathbf{m}^\top \Omega^{-1} \nabla \mathbf{m})^{-1}, \end{aligned}$$

which is positive definite since  $(I - \mathbf{W}^{-1/2} \nabla \mathbf{m} (\nabla \mathbf{m}^\top \mathbf{W}^{-1} \nabla \mathbf{m})^{-1} \nabla \mathbf{m}^\top \mathbf{W}^{-1/2})$  is an idempotent matrix with eigenvalues being either 1 or 0. Let  $\lambda_{\min}(\cdot)$  denote a minimum eigenvalue of a matrix, we then have

$$\lambda_{\min} \left( (\nabla \mathbf{m}^\top \Omega^{-1} \nabla \mathbf{m})^{-1} (\nabla \mathbf{m}^\top \Omega^{-1} \mathbf{W} \Omega^{-1} \nabla \mathbf{m}) (\nabla \mathbf{m}^\top \Omega^{-1} \nabla \mathbf{m})^{-1} - (\nabla \mathbf{m}^\top \mathbf{W}^{-1} \nabla \mathbf{m})^{-1} \right) = c > 0,$$

where the inequality is due to the fact that  $\Omega \neq \mathbf{W}$ .

Second, let

$$\begin{aligned} c_{\min} &= \lambda_{\min} \left( (\nabla \mathbf{m}^\top \Omega^{-1} \nabla \mathbf{m})^{-1} (\nabla \mathbf{m}^\top \Omega^{-1} B \Omega^{-1} \nabla \mathbf{m}) (\nabla \mathbf{m}^\top \Omega^{-1} \nabla \mathbf{m})^{-1} \right. \\ & \left. - (\nabla \mathbf{m}^\top \mathbf{W}^{-1} \nabla \mathbf{m})^{-1} (\nabla \mathbf{m}^\top \mathbf{W}^{-1} B \mathbf{W}^{-1} \nabla \mathbf{m}) (\nabla \mathbf{m}^\top \mathbf{W}^{-1} \nabla \mathbf{m})^{-1} \right). \end{aligned}$$

Note that  $c_{\min}$  can be either positive or negative.

By Weyl's inequality, we have

$$\lambda_{\min} (\Sigma_A^{-1} \Sigma_B \Sigma_A^{-1} - \Sigma_{G,A}^{-1} \Sigma_{G,B} \Sigma_{G,A}^{-1}) \geq c + \delta c_{\min}.$$

Therefore, to ensure the superior property of our GEE estimator (i.e.,  $\Sigma_A^{-1} \Sigma_B \Sigma_A^{-1} - \Sigma_{G,A}^{-1} \Sigma_{G,B} \Sigma_{G,A}^{-1}$  to be positive definite), we can set

$$\delta |c_{\min}| < c, \quad (13)$$

which corresponds to a moderate misspecification of the working variance-covariance matrix  $\mathbf{W}$ .

### 3. Estimating nonlinear models with spatial error: two examples

In this section, we give two examples to show how discrete data can contain the spatially correlated error term and how to use a GEE procedure to estimate the nonlinear models. The first example is for count data, and the second one is for binary response data.

#### 3.1. Example 1: count data with a multiplicative spatial error

A count variable is a variable that takes nonnegative integer values, such as the number of patents applied for by a firm in a given year (e.g., Bloom, Schankerman, and Van Reenen, 2013) and the number of children in the family (e.g., Wooldridge, 2010).

##### 3.1.1. Poisson model

We first model the count data with a conditional Poisson density,  $f(y|\mathbf{x}) = \exp[-\mu] \mu^y / y!$ , where  $y! = 1 \cdot 2 \cdot \dots \cdot (y-1) \cdot y$  and  $0! = 1$ . Denote  $\mu$  as the conditional mean of  $y$ . The Poisson QMLE only requires the conditional mean to be correctly specified. A default assumption for the Poisson distribution is that the mean is equal to the variance. Note that even if  $y_i$  does not follow a Poisson distribution, the QMLE approach will give a consistent estimator if the Poisson density function is used with a correctly specified conditional mean (Gourieroux, Monfort, and Trognon, 1984). Furthermore,  $y_i$  does not even have to be a count variable.

A mean function commonly adopted in applied work is the exponential function:

$$E(y_i | \mathbf{x}_i) = \exp(\mathbf{x}_i \beta_0). \quad (14)$$

When spatial correlations exist, we can characterize the count data model with a multiplicative spatial error. Silva and Tenreiro (2006) consider a Poisson QMLE-type model with multiplicative error terms.

They indicate that the OLS is inconsistent due to the multiplicative error. Now we study a similar model with spatial correlation, i.e.,

$$E(y_i | \mathbf{x}_i, \xi_i) = \xi_i \exp(\mathbf{x}_i \beta_0), \quad (15)$$

where  $\xi_i$  is the multiplicative spatial error term, and we assume the model is characterized by the following features:

- (1)  $\{(\mathbf{x}_i, \xi_i), i = 1, 2, \dots, n\}$  is a mixing sequence on the sampling space  $D_n$ , with a mixing coefficient  $\alpha$ .
- (2)  $E(y_i | \mathbf{x}_i, \xi_i) = \xi_i \exp(\mathbf{x}_i \beta_0)$ .
- (3) For  $i \neq j$ ,  $y_i$  and  $y_j$  are independent conditional on  $\mathbf{x}_i, \mathbf{x}_j, \xi_i$ , and  $\xi_j$ .
- (4)  $\xi_i$  has a conditional multivariate distribution, i.e.,  $E(\xi_i | \mathbf{x}_i) = 1$ ,  $\text{Var}(\xi_i | \mathbf{x}_i) = \tau^2$ , and  $\text{Cov}(\xi_i, \xi_j | \mathbf{x}_i, \mathbf{x}_j) = \tau^2 \cdot c(d_{ij}, \rho)$ , where  $c(d_{ij}, \rho)$  is the correlation function of  $\xi_i$  and  $\xi_j$ , and  $\rho$  is the parameter.

To be more explicit, the log conditional mean is assumed to be

$$\log E(y_i | \mathbf{x}_i, \xi_i) = \mathbf{x}_i \beta + \log \xi_i. \quad (16)$$

Also, we can write  $\lambda_i = \exp(\mathbf{x}_i \beta + \log \xi_i)$ . The assumed conditional probability mass function is

$$\mathbb{P}(y_i = y | \xi_i, \mathbf{x}_i^\top) = \exp(-\lambda_i) \lambda_i^y / y!. \quad (17)$$

Under the above assumptions, we can integrate out  $\xi_i$  by using the law of iterated expectations,

$$E(y_i | \mathbf{x}_i, D_n) = E(E(y_i | \mathbf{x}_i, \xi_i) | \mathbf{x}_i, D_n) = \exp(\mathbf{x}_i \beta_0), \quad (18)$$

where we suppress the condition on  $D_n$  in the last equality. The QMLE gives a consistent estimator for the mean parameters, which solves

$$\check{\beta}_{\text{QMLE}} = \text{argmax}_{\beta} \sum_{i=1}^n l_i(\beta) = \sum_{i=1}^n y_i \mathbf{x}_i \beta - \sum_{i=1}^n \exp(\mathbf{x}_i \beta) - \sum_{i=1}^n \log(y_i!). \quad (19)$$

Its score function is

$$\sum_{i=1}^n \mathbf{x}_i^\top \left[ y_i - \exp(\mathbf{x}_i \check{\beta}_{\text{QMLE}}) \right] = \mathbf{0}. \quad (20)$$

Since the above estimator does not account for any heteroskedasticity or spatial correlation, a robust estimator for the asymptotic variance of the QMLE estimator is provided as follows:

$$\widehat{\text{Avar}}(\check{\beta}_{\text{QMLE}}) = \left[ \sum_{i=1}^n \exp(-\mathbf{x}_i \check{\beta}_{\text{QMLE}}) \mathbf{x}_i^\top \mathbf{x}_i \right]^{-1} \sum_{i=1}^n \sum_{j=1}^n k(d_{ij}) \mathbf{x}_i^\top \check{u}_i \check{u}_j \mathbf{x}_j \left[ \sum_{i=1}^n \exp(-\mathbf{x}_i \check{\beta}_{\text{QMLE}}) \mathbf{x}_i^\top \mathbf{x}_i \right]^{-1},$$

where  $k(d_{ij})$  is a kernel function depending on the distance between observations  $i$  and  $j$ , and  $\check{u}_i = y_i - \exp(\mathbf{x}_i \check{\beta}_{\text{QMLE}})$ .

Moreover, a very specific aspect of the Poisson distribution is that we can write down the conditional variances of  $y$  as

$$\text{Var}(y_i | \mathbf{x}_i, D_n) = \exp(\mathbf{x}_i \beta_0) + \exp(2\mathbf{x}_i \beta_0) \cdot \tau^2. \quad (22)$$

The conditional variance of  $y_i$  given  $\mathbf{x}_i$  is a function of both the level and the quadratic of the conditional mean. The conditional Poisson distribution is characterized by the equality between its conditional variance and conditional mean, i.e.,  $\text{Var}(y_i | \mathbf{x}_i) = \exp(\mathbf{x}_i \beta_0)$ . One can relax the variance assumption

to  $\text{Var}(y_i|\mathbf{x}_i) = \sigma^2 \exp(\mathbf{x}_i\beta_0)$  with an overdispersion or underdispersion constant parameter  $\sigma^2$ . Obviously, there is an over-dispersion in (22) since  $\exp(2\mathbf{x}_i\beta_0) \cdot \tau^2 \geq 0$ , and the over-dispersion parameter is  $1 + \exp(\mathbf{x}_i\beta_0) \cdot \tau^2$ , which is changing with  $\mathbf{x}_i$ . This variance structure does not coincide with conditional Poisson distribution. Moreover, the conditional covariances can be written in the following form:

$$\text{Cov}(y_i, y_j|\mathbf{x}_i, \mathbf{x}_j, D_n) = \exp(\mathbf{x}_i\beta_0) \exp(\mathbf{x}_j\beta_0) \cdot \tau^2 \cdot c(d_{ij}, \rho). \quad (23)$$

In the group-level notation,

$$\text{E}(\mathbf{y}_g|\mathbf{x}_g, D_G) = \exp(\mathbf{x}_g\beta_0). \quad (24)$$

Let  $\mathbf{W}_g$  be the variance-covariance matrix for the group  $g$  evaluated at the true value  $\beta_0, \rho$ . The variance of the  $l$ -th element in the group  $g$  is

$$v_{gl} = \exp(\mathbf{x}_{gl}\beta_0) (1 + \exp(\mathbf{x}_{gl}\beta_0) \cdot \tau^2), \quad (25)$$

and the covariance of the  $l$ th and  $m$ th elements in group  $g$  is

$$r_{glm} = \exp(\mathbf{x}_{gl}\beta_0) \exp(\mathbf{x}_{gm}\beta_0) \cdot \tau^2 \cdot c(d_{glm}, \rho). \quad (26)$$

Here  $\gamma = (\tau^2, \rho)^\top$ , and let  $\check{\gamma} = (\check{\tau}^2, \check{\rho})^\top$  be an estimator for  $\gamma$ . Let  $\check{\beta}_{\text{QMLE}}$  be the QMLE estimator in the first step. Then the elements in  $\mathbf{W}_g$  can be estimated as:

$$\hat{v}_{gl} = \exp(\mathbf{x}_{gl}\check{\beta}_{\text{QMLE}}) + \exp(2\mathbf{x}_{gl}\check{\beta}_{\text{QMLE}}) \cdot \check{\tau}^2, \quad (27)$$

and

$$\hat{r}_{glm} = \exp(\mathbf{x}_{gl}\check{\beta}_{\text{QMLE}}) \exp(\mathbf{x}_{gm}\check{\beta}_{\text{QMLE}}) \cdot \check{\tau}^2 \cdot c(d_{glm}, \check{\rho}), \quad (28)$$

where  $d_{glm}$  is the distance between the object  $l$  and  $m$  in the group  $g$ , and it corresponds to the distance  $d_{ij}$  with the label  $i, j$ .

### 3.1.2. Negative binomial II model

Now we discuss the Negative Binomial Model for count data. Since the conditional variances and covariances can be written in a specific form, we would consider the NegBin II model (NBII hereafter) of Cameron and Trivedi (1986) as an appropriate choice. The NBII model can be derived from a Poisson model with multiplicative error. With an exponential mean, we assume

$$y_i|\mathbf{x}_i, \xi_i, \varepsilon_i, D_n \sim \text{Poisson}[\varepsilon_i\xi_i \exp(\mathbf{x}_i\beta_0)] \text{ with } \xi_i > 0, \varepsilon_i > 0,$$

where  $\xi_i$  follows from condition (4) in Section 3.1.1, and  $\varepsilon_i$  follows from Gamma distribution with the density:

$$\frac{\psi^{-\psi}}{\Gamma(\psi)} \varepsilon_i^{\psi-1} \exp(-\varepsilon_i\psi), \quad (29)$$

where  $\text{E}(\varepsilon_i) = 1$ ,  $\text{Var}(\varepsilon_i) = 1/\psi$ , for  $\psi > 0$ , and  $\Gamma(\cdot)$  is the gamma function. Moreover, we assume that  $\varepsilon_i$ s are independent of  $\xi_i$ s and  $\mathbf{x}_i$ s, and  $\varepsilon_i$ s are i.i.d. We also assume that for  $i \neq j$ ,  $y_i$  and  $y_j$  are independent conditional on  $\mathbf{x}_i, \mathbf{x}_j, \xi_i, \xi_j, \varepsilon_i$ , and  $\varepsilon_j$ .

Under the above assumptions for the Poisson distribution, with the conditional mean as in (18) and conditional variance similar to (22),  $y_i|\mathbf{x}_i$  is shown to follow a negative binomial II distribution. The model implies overdispersion as well, and the amount of overdispersion increases with the conditional mean,

$$\text{Var}(y_i|\mathbf{x}_i, D_n) = \exp(\mathbf{x}_i\beta_0) (1 + \exp(\mathbf{x}_i\beta_0) \cdot [\tau^2/\psi + \tau^2 + 1/\psi]). \quad (30)$$

The covariance function in Eq. (26) for  $l \neq m$  stays the same. The pre-estimator  $\check{\beta}$  and  $\check{\psi}$  can be obtained by a QMLE maximizing the likelihood function for the standard NBII model. See, e.g., Gouriéroux, Monfort, and Trognon (1984). The estimation of  $\tau$  will be discussed in the following sub-section.

### 3.1.3. GEE estimation

In both cases, let  $\check{\theta} = \check{\beta}_{\text{QMLE}}$  and  $\check{\gamma} = (\check{\tau}, \check{\rho})$  ( $\check{\gamma} = (\check{\psi}, \check{\tau}, \check{\rho})$  for the NBII case). Let  $\mathbf{x}_g$  be  $L \times p$ , and  $\mathbf{x}_g \beta$  be  $L \times 1$ . Moreover,  $\exp(\mathbf{x}_g \beta)$  is a  $L \times 1$  vector, where  $\exp$  is applied elementwise to  $\mathbf{x}_g \beta$ . Based on the conditional distribution, the first-order condition for GEE is

$$\sum_g \mathbf{x}_g^\top \text{diag}(\exp(\mathbf{x}_g \hat{\beta}_{\text{GEE}})) \mathbf{W}_g^{-1}(\check{\gamma}, \check{\theta}) [\mathbf{y}_g - \exp(\mathbf{x}_g \hat{\beta}_{\text{GEE}})] = 0. \quad (31)$$

$\hat{\beta}_{\text{GEE}}$  is consistent and follows a normal distribution asymptotically by Theorem 1 and 2 in Sections 4.3 and 4.4. We will abbreviate  $\mathbf{W}_g^{-1}(\check{\gamma}, \check{\theta})$  to  $\hat{\mathbf{W}}_g^{-1}$  in the following text. We denote  $\hat{\mu}_g = \exp(\mathbf{x}_g \hat{\beta}_{\text{GEE}})$  and  $\hat{\mathbf{u}}_g = \mathbf{y}_g - \exp(\mathbf{x}_g \hat{\beta}_{\text{GEE}})$ . Following the spatial heteroscedasticity and autocorrelation (HAC) consistent estimation literature (see, e.g., Kelejian and Prucha, 1999), the asymptotic variance estimator could be constructed as follows:

$$\begin{aligned} \widehat{\text{Avar}}(\hat{\beta}_{\text{GEE}}) &= \left( \sum_g \mathbf{x}_g^\top \text{diag}(\hat{\mu}_g) \hat{\mathbf{W}}_g^{-1} \text{diag}(\hat{\mu}_g) \mathbf{x}_g \right)^{-1} \\ &\quad \left( \sum_g \sum_h k(d_{gh}) \mathbf{x}_g^\top \text{diag}(\hat{\mu}_g) \hat{\mathbf{W}}_g^{-1} \hat{\mathbf{u}}_g \hat{\mathbf{u}}_h^\top \hat{\mathbf{W}}_h^{-1} \text{diag}(\hat{\mu}_h) \mathbf{x}_h \right) \\ &\quad \left( \sum_g \mathbf{x}_g^\top \text{diag}(\hat{\mu}_g) \hat{\mathbf{W}}_g^{-1} \text{diag}(\hat{\mu}_g) \mathbf{x}_g \right)^{-1}, \end{aligned}$$

where  $k(d_{gh})$  is a kernel function depending on the distances between groups.

The parameters,  $\tau^2$  and  $\rho$ , can be estimated using the Poisson QMLE residuals. Let  $\check{u}_i^2 = [y_i - \exp(\mathbf{x}_i \check{\beta}_{\text{QMLE}})]^2$  be the squared residuals from the Poisson QMLE. Based on Eq. (25),  $\tau^2$  ( $\tau^2/\psi$  for the NBII case) can be estimated as the coefficient by regressing  $\check{u}_i^2 - \exp(\mathbf{x}_i \check{\beta}_{\text{QMLE}})$  on  $\exp(2\mathbf{x}_i \check{\beta}_{\text{QMLE}})$ . The estimation of  $\rho$  depends on the specific form of  $c(d_{ij}, \rho)$ . We can impose, though perhaps wrongfully, a structure on the true covariance. For example, suppose that the covariance structure of  $e_i$  and  $e_j$  is  $\exp\left(\frac{\rho}{d_{ij}}\right) - 1$ , then an estimator for  $\rho$  is

$$\hat{\rho} = \mathbf{argmin}_\rho \sum_{i=1}^n \sum_{j=1}^n \left\{ \frac{\check{u}_i \check{u}_j}{\exp(\mathbf{x}_i \check{\beta}) \exp(\mathbf{x}_j \check{\beta})} - \left[ \exp\left(\frac{\rho}{d_{ij}}\right) - 1 \right] \right\}^2. \quad (33)$$

Then  $\hat{\mathbf{W}}_g$  is obtained by plugging  $\hat{\tau}^2$  and  $\hat{\rho}$  back into the variance-covariance matrix. We can also directly calculate  $\hat{\rho}$  as

$$\hat{\rho} = \frac{1}{n \cdot (n-1)} \sum_{i=1}^n \sum_{j \neq i}^n \left[ \log \left( \frac{\check{u}_i \check{u}_j}{\exp(\mathbf{x}_i \check{\beta}) \exp(\mathbf{x}_j \check{\beta})} + 1 \right) \cdot d_{ij} \right]. \quad (34)$$

Then we can specify a GEE working correlation matrix to estimate (31).

### 3.2. Example 2: binary response data with spatial correlation in the latent error

We start from the Probit model:

$$y_i = 1 [y_i^* > 0], \quad (35)$$

$$y_i^* = \mathbf{x}_i \boldsymbol{\beta} + e_i, \quad (36)$$

where  $y_i^*$  is an unobservable latent variable. Now let  $\mathbf{e} = (e_1, e_2, \dots, e_n)^\top$  be the vector of latent spatially correlated error. For example, Pinkse and Slade (1998) use the following assumption of  $\mathbf{e}$ :

$$\mathbf{e} = \rho M_w \mathbf{e} + \boldsymbol{\varepsilon}, \quad (37)$$

where  $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$  is a vector of independent standard normal distribution.  $M_w$  is an  $n \times n$  weighting matrix, where its diagonal elements are zeroes, and its off-diagonal element, for example  $M_{w,ij}$  with  $i \neq j$ , is inversely proportional to the distances between location  $i$  and  $j$ .  $\rho$  is a spatial correlation parameter. In this case,  $\mathbf{e}$  can be written as a function of  $\boldsymbol{\varepsilon}$ ,

$$\mathbf{e} = (I - \rho M_w)^{-1} \boldsymbol{\varepsilon}. \quad (38)$$

Thus, the conditional expectation of  $\mathbf{e}$  is zero. The variance-covariance matrix of  $\mathbf{e}$  is

$$\text{Var}(\mathbf{e}|\mathbf{x}, D_n) = (I - \rho M_w)^{-1} (I - \rho M_w)^{-1\top}. \quad (39)$$

If we assume that  $\mathbf{e}|\mathbf{x}$  has a multivariate normal distribution with zero mean and a variance matrix specified in (39), the conditional mean is correctly specified as follows:

$$E(y_i|\mathbf{x}_i, D_n) = \Phi_i(\mathbf{x}_i \boldsymbol{\beta}), \quad (40)$$

where  $\Phi_i$  is the marginal normal distribution function with its variance being the  $i$ -th element of the diagonal of (39). Let  $\phi_i(\cdot)$  denote the corresponding density function. Moreover, the conditional variance function for a Bernoulli distribution as

$$\text{Var}(y_i|\mathbf{x}_i, D_n) = \Phi_i(\mathbf{x}_i \boldsymbol{\beta}) [1 - \Phi_i(\mathbf{x}_i \boldsymbol{\beta})]. \quad (41)$$

Taking the Bernoulli QMLE as an example, which is obtained by maximizing the Probit log-likelihood. The log-likelihood function for each observation is

$$l_i(\boldsymbol{\beta}) = y_i \log \Phi_i(\mathbf{x}_i \boldsymbol{\beta}) + (1 - y_i) \log [1 - \Phi_i(\mathbf{x}_i \boldsymbol{\beta})]. \quad (42)$$

Let  $\check{u}_i = y_i - \Phi_i(\mathbf{x}_i \check{\boldsymbol{\beta}})$  (for  $i = 1, 2, \dots, n$ ) be the residual from the QMLE estimation. At this stage, an estimator for the asymptotic variance of  $\check{\boldsymbol{\beta}}_{\text{QMLE}}$  can be computed as follows:

$$\begin{aligned} \widehat{\text{Avar}}(\check{\boldsymbol{\beta}}_{\text{QMLE}}) &= \left( \sum_{i=1}^n \frac{\phi_i^2(\mathbf{x}_i \check{\boldsymbol{\beta}}_{\text{QMLE}}) \mathbf{x}_i^\top \mathbf{x}_i}{\Phi_i(\mathbf{x}_i \check{\boldsymbol{\beta}}_{\text{QMLE}}) [1 - \Phi_i(\mathbf{x}_i \check{\boldsymbol{\beta}}_{\text{QMLE}})]} \right)^{-1} \\ &\quad \left( \sum_{i=1}^n \sum_{j=1}^n k(d_{ij}) \frac{\phi_i(\mathbf{x}_i \check{\boldsymbol{\beta}}_{\text{QMLE}}) \phi_j(\mathbf{x}_j \check{\boldsymbol{\beta}}_{\text{QMLE}}) \mathbf{x}_i^\top \check{u}_i \check{u}_j \mathbf{x}_j}{\Phi_i(\mathbf{x}_i \check{\boldsymbol{\beta}}_{\text{QMLE}}) [1 - \Phi_i(\mathbf{x}_i \check{\boldsymbol{\beta}}_{\text{QMLE}})]} \right) \\ &\quad \left( \sum_{i=1}^n \frac{\phi_i^2(\mathbf{x}_i \check{\boldsymbol{\beta}}_{\text{QMLE}}) \mathbf{x}_i^\top \mathbf{x}_i}{\Phi_i(\mathbf{x}_i \check{\boldsymbol{\beta}}_{\text{QMLE}}) [1 - \Phi_i(\mathbf{x}_i \check{\boldsymbol{\beta}}_{\text{QMLE}})]} \right)^{-1}, \end{aligned}$$

where  $k(d_{ij})$  is the kernel weight function that depends on pairwise distances. This QMLE and its robust variance-covariance estimator provide a legitimate way of estimating the spatial Probit model.

We use QMLE as a first-step estimator. An estimator for the working variance matrix for each group is

$$\check{v}_{gl} = \Phi_{gl}(\mathbf{x}_{gl} \check{\boldsymbol{\beta}}_{\text{QMLE}}) \left[ 1 - \Phi_{gl}(\mathbf{x}_{gl} \check{\boldsymbol{\beta}}_{\text{QMLE}}) \right]. \quad (44)$$

Furthermore, we assume the working correlation function for  $l$ -th and  $m$ -th elements in group  $g$  is

$$\pi_{glm} = C(d_{glm}, \rho). \quad (45)$$

For example,

$$C(d_{glm}, \rho) = \frac{\rho}{d_{glm}} \text{ or } \exp\left(-\frac{d_{glm}}{\rho}\right). \quad (46)$$

Let  $\check{u}_{gl}$  be the QMLE residual for  $l$ -th element in  $r$ -th group, and  $\hat{r}_{gl} = \check{u}_{gl}/\sqrt{\check{v}_{gl}}$  be the standardized residual. Using the correlations within groups, one estimator of  $\rho$  is

$$\hat{\rho} = \mathbf{argmin}_{\rho} \sum_g \sum_{l=1}^L \sum_{m<l} [\hat{r}_{gl}\hat{r}_{gm} - C(d_{glm}, \rho)]^2, \quad (47)$$

for  $l < m$ .

Define  $\Phi_g(\mathbf{x}_g \hat{\beta}_{GEE}) = (\Phi_{g1}(\mathbf{x}_{g1} \hat{\beta}_{GEE}), \dots, \Phi_{gL}(\mathbf{x}_{gL} \hat{\beta}_{GEE}))'$ , and  $\hat{\mu}_g = (\phi_{g1}, \dots, \phi_{gL})$  with  $\phi_{gl} = \phi(\mathbf{x}_{gl} \hat{\beta})$ ,  $l = 1, 2, \dots, L$ . Similarly, the second-step GEE estimator solves

$$\sum_g \mathbf{x}_g^\top \text{diag}\{\phi_{gl}\} \hat{\mathbf{W}}_g^{-1} (\mathbf{y}_g - \Phi_g(\mathbf{x}_g \hat{\beta}_{GEE})) = \mathbf{0}. \quad (48)$$

$\hat{\beta}_{GEE}$  is consistent and follows a normal distribution asymptotically by [Theorems 1 and 2](#) of [Sections 4.3 and 4.4](#).  $\hat{\beta}_{GEE}$  is consistent even for misspecified spatial correlation structure  $\hat{\mathbf{W}}_g$ . We could again construct the spatial HAC-type variance estimator as follows

$$\begin{aligned} \widehat{\text{Avar}}(\hat{\beta}_{GEE}) &= \left( \sum_g \mathbf{x}_g^\top \text{diag}\{\hat{\mu}_g\} \hat{\mathbf{W}}_g^{-1} \text{diag}\{\hat{\mu}_g\} \mathbf{x}_g \right)^{-1} \\ &\quad \left( \sum_g \sum_h k(d_{gh}) \mathbf{x}_g^\top \text{diag}\{\hat{\mu}_g\} \hat{\mathbf{W}}_g^{-1} \hat{\mathbf{u}}_g \hat{\mathbf{u}}_h^\top \hat{\mathbf{W}}_h^{-1} \text{diag}\{\hat{\mu}_h\} \mathbf{x}_h \right) \\ &\quad \left( \sum_g \mathbf{x}_g^\top \text{diag}\{\hat{\mu}_g\} \hat{\mathbf{W}}_g^{-1} \text{diag}\{\hat{\mu}_g\} \mathbf{x}_g \right)^{-1}, \end{aligned}$$

where  $k(d_{gh})$  is a kernel function that depends on the distances between groups.

## 4. Theorems

In this section, we present the assumptions and investigate the theoretical properties of our GEE estimation. In [Sections 4.1 and 4.2](#), we introduce some notations and definitions, while [Sections 4.3 and 4.4](#) present the consistency and asymptotic normality of the GEE estimator in (7); [Section 4.5](#) demonstrates the consistency of the estimation of the variance-covariance matrix of the proposed GEE estimator.

### 4.1. Notations

We need ULLNs and CLTs for analyzing the properties of our proposed GEE estimator. While theories for temporal dependence data have been well established in the literature (see, e.g., [Davidson, 2021](#)), they are not suitable for our analysis since we are working with spatial data that lacks a natural order. Using a distance measure defined based on the maximum metric, [Jenish and Prucha \(2012\)](#) develop a ULLN and a CLT for  $\alpha$ -mixing random fields on unevenly spaced lattices that allow nonstationary processes

with trending moments. However, the mixing property can fail for quite a few reasons. Thus, we adopt the notion of NED as in Jenish and Prucha (2012) which refers to a generalized class of random fields that is "closed with respect to infinite transformations."

Let  $Z = \{Z_{n,i}, i \in D_n, n \geq 1\}$  and  $\varepsilon = \{\varepsilon_{n,i}, i \in T_n, n \geq 1\}$  be triangular arrays of random fields defined on a probability space  $(\Omega_\varepsilon, \mathcal{F}, P)$ , where  $T_n$  is a larger lattice with  $D_n \subseteq T_n \subseteq D$ .  $D$  satisfies the following Assumption A.1. The cardinality of  $D_n$  and  $T_n$  satisfies  $\lim_{n \rightarrow \infty} |D_n| \rightarrow \infty$ ,  $\lim_{n \rightarrow \infty} |T_n| \rightarrow \infty$ . For  $i, j \in \mathbb{R}^d$ ,

we consider a metric  $d(i, j) \stackrel{\text{def}}{=} \max_{1 \leq l \leq d} |j_l - i_l|$  with the norm  $\|i\|_\infty = \max_{1 \leq l \leq d} |i_l|$ , where  $i_l$  is the  $l$ -th component of  $i$ . The distance between any subsets  $U, V \in D$  is defined as  $d(U, V) = \inf\{d(i, j) : i \in U \text{ and } j \in V\}$ . For a vector (matrix)  $A$ , let  $\|A\|_2$  denote its  $L_2$ -norm, and  $\|A\|_a$  return a vector (matrix) wherein each element is the absolute value of the corresponding element in vector (matrix)  $A$ . For any random vector  $X$ , let  $\|X_{n,i}\|_p = (E |X_{n,i}|^p)^{1/p}$  denote its  $L_p$ -norm, given that the absolute  $p$ -th moment exists. In the case of  $p = 2$ , we abbreviate  $\|X_{n,i}\|_2$  as  $\|X_{n,i}\|$ . Finally, We let  $\mathcal{F}_{n,i}(s) = \sigma(\varepsilon_{n,j} : j \in D_n, d(i, j) \leq s)$  be the  $\sigma$ -field generated by random vectors  $\varepsilon_{n,j}$  located within a distance of  $s$  from  $i$ .

## 4.2. Definitions

We start with definitions needed for the consistency and asymptotic normality of our estimator.

**Definition 1.** Let  $Z = \{Z_{n,i}, i \in D_n, n \geq 1\}$  and  $\varepsilon = \{\varepsilon_{n,i}, i \in D_n, n \geq 1\}$  be random fields with  $\|Z_{n,i}\|_p < \infty, p \geq 1$ , where  $D_n \subseteq D$ , and its cardinality is  $|D_n| = n$ . Let  $\{d_{n,i}, i \in D_n, n \geq 1\}$  be an array of finite positive constants. Then the random field  $Z$  is said to be  $L_p$ -near-epoch dependent on the random field  $\varepsilon$  ( $L_p$ -NED on  $\varepsilon$ ) if

$$\|Z_{n,i} - E(Z_{n,i} | \mathcal{F}_{n,i}(s))\|_p < d_{n,i} \varphi(s)$$

holds for some sequence  $\varphi(s) \geq 0$  with  $\lim_{s \rightarrow \infty} \varphi(s) = 0$ , where  $\varphi(s)$  denotes the NED coefficient, and  $d_{n,i}$  is a NED scaling factor. Furthermore, if  $\psi(s) = s^{-\mu}$  for some  $\mu > \lambda > 0$ , then  $Z$  is referred to as  $L_p$ -NED on  $\varepsilon$  of size  $-\lambda$ . Moreover, if  $\sup_n \sup_{i \in D_n} d_{n,i} < \infty$ , then  $Z$  is called uniformly  $L_p$ -NED on  $\varepsilon$ .

We will present the  $L_2$ -NED properties of a random field  $Z$  on some  $\alpha$ -mixing random field  $\varepsilon$ . The definition of the  $\alpha$ -mixing coefficient employed in the article is stated as follows.

**Definition 2.** Let  $\mathcal{A}$  and  $\mathcal{B}$  be two  $\sigma$ -algebras of  $\mathcal{F}$ , and let

$$\alpha(\mathcal{A}, \mathcal{B}) = \sup_{A, B} |P(A \cap B) - P(A)P(B)|, A \in \mathcal{A}, B \in \mathcal{B},$$

For  $U \subseteq D_n$  and  $V \subseteq D_n$ , let  $\sigma_n(U) = \sigma(\varepsilon_{n,i}, i \in U)$  ( $\sigma_n(V) = \sigma(\varepsilon_{n,i}, i \in V)$ ) and  $\alpha_n(U, V) = \alpha(\sigma_n(U), \sigma_n(V))$ . Then, the  $\alpha$ -mixing coefficients for the random field  $\varepsilon$  are defined as:

$$\bar{\alpha}(u, v, h) = \sup_n \sup_{U, V} (\alpha_n(U, V), |U| \leq u, |V| \leq v, d(U, V) \geq h).$$

where  $d(\cdot, \cdot)$  is the distance measure based on the maximum metric.

Compared with the one applied in temporal analysis, the above-defined  $\alpha$ -mixing also depends on the size of the subsets, as given a fixed distance in random fields, it is natural to expect more dependence between two larger sets than between two smaller sets. Moving forward, we will suppress the dependence on  $n$  for a triangular array if there is no confusion in the context.

## 4.3. Consistency

In the following, we present assumptions needed for establishing asymptotic theories.



**Assumption A.1.** (Sampling point). The lattice  $D \subseteq \mathbb{R}^d$ ,  $d \geq 1$ , is infinitely countable. The distance  $d(i, j)$  between any two different individual units  $i$  and  $j$  in  $D$  is at least larger than a positive constant, i.e.,  $\forall i, j \in D : d(i, j) \geq \rho_0$ . W.l.o.g., we assume  $\rho_0 > 1$ .

This is the basic assumption on the distance measure and the lattices. We do not consider infill asymptotic framework and therefore we impose a minimum distance assumption between observations.

**Assumption A.2.** (Decay dependence).  $\{y_i\}$  is  $L_4$ - uniformly NED on the  $\alpha$ - mixing random field  $\varepsilon = \{\varepsilon_i, i \in D_n\}$ , where  $\varepsilon_i = (x_i, u_i)$  ( $u_i$ 's are some underlying innovation processes). For the  $\alpha$ -mixing coefficient  $\bar{\alpha}$ , it holds that  $\bar{\alpha}(u, v, r) \leq (u + v)^\tau \hat{\alpha}(r)$  for a constant  $\tau \geq 0$  and a function  $\hat{\alpha}(\cdot)$  that satisfies  $\lim_{r \rightarrow \infty} \hat{\alpha}(r) = 0$  and  $\sum_{r=1}^{\infty} r^{d-1} \hat{\alpha}(r) < \infty$ . The NED constant is denoted by  $d_{n,i}$ , which satisfies  $\sup_{n,i \in T_n} d_{n,i} < \infty$ . The NED coefficient is denoted by  $\psi(s)$ , which satisfies  $\lim_{s \rightarrow \infty} \psi(s) = 0$ , and  $\sum_{r=0}^{\infty} r^{d-1} \psi(r) < \infty$ .

**Assumption A.3.** (Parameter space). The parameter space  $\Theta \times \Gamma$  is a compact subset of  $\mathbb{R}^{p+q}$  with euclidean metric  $\|\cdot\|_2$ .  $q_g(\theta, \gamma)$ ,  $s_g(\theta, \gamma)$ , and  $h_g(\theta, \gamma)$  defined in Section 2.2 are functions from  $\Theta \times \Gamma$  to  $\mathbb{R}^1$ ,  $\mathbb{R}^p$ , and  $\mathbb{R}^{p^2}$ , respectively. These functions are measurable for each  $\theta \in \Theta$  and  $\gamma \in \Gamma$ , and are Lipschitz continuous on  $\Theta \times \Gamma$ .

From now on, we work with group-level asymptotics. We define the field  $\tilde{\varepsilon} = \{\varepsilon_g : g \in 1, \dots, G\}$  with grouped observations. First of all, suppose that  $D_n$  is divided into  $G$  blocks with  $\cup_1^G B_g = D_n \subset T_n$  (recall that the group-level lattice is denoted as  $D_G$ ). Define the distance between two groups,  $g$  and  $h$ , as  $d(g, h) = \min_{i \in B_g, j \in B_h} d(i, j)$ . Then, the  $\alpha$ -mixing coefficient between group  $U = \{g_1, \dots, g_L\}$  and group  $V = \{h_1, \dots, h_M\}$  is thus  $\tilde{\alpha}(u, v, r) = \sup_{L \leq u, M \leq v, d(U, V) \geq r} \alpha(\sigma(U), \sigma(V))$ , where  $d(U, V) = \min_{l \in 1 \dots L, m \in 1, \dots, M} d(g_l, h_m)$ . As  $L$  is assumed to be fixed, the grouping observations  $\tilde{\varepsilon}$  will satisfy the mixing coefficients restrictions imposed in Assumption A.2. Moreover, since  $L$  is the same for every group, we shall have  $\tilde{\alpha}(u, v, r) = (uL + vL)^\tau \hat{\alpha}(r)$ .

**Assumption A.4.** (Moment assumptions).  $E \sup_{\theta \in \Theta} |m_{g,i}|^r \leq C_1$ ,  $E \sup_{\theta \in \Theta, \gamma \in \Gamma} |w_{g,ij}|^r \leq C_2$ ,  $E |y_{g,i}|^r \leq C_3$ , and  $E \sup_{\theta \in \Theta} |\nabla_\theta m_{g,i}|^r \leq C_4$ , where  $C_1, C_2, C_3$ , and  $C_4$  are positive constants, and  $w_{g,ij}$ ,  $y_{g,i}$ , and  $m_{g,i}$  are the element-wise components for  $\mathbf{W}_g^{-1}(\theta, \gamma)$ ,  $\mathbf{y}_g$ , and  $\mathbf{m}_g(\theta, \gamma)$ .  $m_{g,i}$  and  $w_{g,ij}$  are continuously differentiable up to the third order derivatives, and their  $r$ -th moments (after taking the supreme over the parameter space) are bounded up to the second order derivatives, with  $r > 4p'$  and  $p' \geq 1$ .

**Remark 2.** The moment conditions in Assumption A.4 are needed for establishing the ULLN to show the consistency of our proposed estimator; see the proof for Lemma 1 in supplementary material. These conditions are primitive conditions and widely employed in the literature; see, e.g., Lemma 2.5 in Newey and McFadden (1994) or Assumption 3 in Newey and Powell (2003).

**Assumption A.5.** (Identifiability). The true parameter  $\theta_0$  is the unique minimizer for the objective function in the sense that, for any  $\varepsilon > 0$ , there exists a positive number  $c_0$  such that  $\liminf_{G \rightarrow \infty} \inf_{\theta \in \Theta : \|\theta - \theta_0\|_2 \geq \varepsilon} Q_G(\check{\theta}, \check{\gamma}, \theta) > c_0 + Q_\infty(\check{\theta}, \check{\gamma}, \theta_0)$ .

Assumption A.2 concerns the  $L_2$ -NED property of data-generating processes. See Section S2 in supplementary material for detailed verification of the special cases. It should be noted that by Lyapunov inequality, if  $\{y_i\}$  is  $L_k$ -NED, it is also  $L_l$ -NED with the same coefficients,  $d_{n,i}$  and  $\psi(s)$ , for any  $l \leq k$ . In fact, as we work with the group-level asymptotics, we can also directly replace the assumption by the NED property of  $\mathbf{y}_g$  on  $\{\varepsilon_g\}$ . This assumption also imposes the algebraic decaying rate of an underlying process, and will not restrict the dependence within groups as we set the group size

to be fixed. Assumption A.3 contains the standard regularities assumptions. Assumption A.4 is a collection of moment conditions on the statistical objects involved in the estimation and also on each element of Hessian matrices, ensuring the boundedness of the moments. Assumption A.5 is a condition on the identification of our proposed estimator. It can be implied by the positive definiteness of  $\mathbf{W}_g^{-1}(\theta, \gamma)$  and the same identification assumption  $\liminf_{G \rightarrow \infty} \inf_{\theta \in \Theta: |\theta - \theta^0|_2 \geq \varepsilon} Q'_G(\theta) > c_0 + Q'_\infty(\theta)$  on  $Q'_G(\theta) \stackrel{\text{def}}{=} \frac{1}{|D_G|} \sum_{g \in |D_G|} \mathbb{E} [\mathbf{y}_g - \mathbf{m}_g(\mathbf{x}_g; \theta)]^\top [\mathbf{y}_g - \mathbf{m}_g(\mathbf{x}_g; \theta)]$  and  $Q'_\infty(\theta) \stackrel{\text{def}}{=} \lim_{G \rightarrow \infty} Q'_G(\theta)$ . It can be shown that  $\liminf_{G \rightarrow \infty} \inf_{\theta \in \Theta: |\theta - \theta^0|_2 \geq \varepsilon} Q_G(\check{\theta}, \check{\gamma}, \theta) > \lambda_{\min}\{\mathbf{W}_g^{-1}(\check{\theta}, \check{\gamma})\}(c_0 + Q'_\infty(\theta))$ , where  $\lambda_{\min}\{\mathbf{W}_g^{-1}(\check{\theta}, \check{\gamma})\}$  is the minimum eigenvalue of the matrix  $\mathbf{W}_g^{-1}(\check{\theta}, \check{\gamma})$ . When  $\inf_{\theta \in \Theta, \gamma \in \Gamma} \lambda_{\min}\{\mathbf{W}_g^{-1}(\theta, \gamma)\} > c$  for  $c > 0$ , Assumption A.5 is then satisfied with probability approaching one. Given these assumptions, we can provide the consistency property of our estimation.

**Theorem 1.** (Consistency). *Under Assumptions A.1–A.5, the proposed GEE estimator obtained by solving Eq. (7) is consistent:  $|\hat{\theta} - \theta^0|_2 \xrightarrow{P} 0$  as  $G \rightarrow \infty$ .*

Theorem 1 indicates that the consistency of  $\hat{\theta}$  does not rely on the consistent estimation for  $\gamma$  in the first step as long as the number of groups tends to infinity and the conditional mean function is correctly specified. The proof is provided in Section S1.2 of supplementary material.

#### 4.4. Normality

To further establish the asymptotic normality of the estimate, we additionally impose the following assumptions:

**Assumption A.6.** (Decaying dependence). *The function  $\hat{\alpha}$  satisfies  $\sum_{r=1}^{\infty} r^{(d\tau^*+d)-1} L^{\tau^*} \hat{\alpha}^{\delta/(2+\delta)}(r) < \infty$  for  $\delta > 0$  and  $\tau^* = \delta\tau/(4 + 2\delta)$ .*

**Assumption A.7.** (Nuisance plug-in). *The limiting points of the nuisance parameters,  $\theta^*$  and  $\gamma^*$ , lie in the interiors of  $\Theta$  and  $\Gamma$ , respectively. Furthermore, we assume  $|\check{\gamma} - \gamma^*|_2 = \mathcal{O}_p(G^{-1/2})$  and  $|\check{\theta} - \theta^*|_2 = \mathcal{O}_p(G^{-1/2})$ .*

Define

$$\begin{aligned} AS_G = & \frac{1}{G} \sum_g \mathbb{E} \left[ \nabla \mathbf{m}_g^\top(\theta^0) \mathbf{W}_g^{-1}(\theta^*, \gamma^*) \mathbf{u}_g \mathbf{u}_g^\top \mathbf{W}_g^{-1}(\theta^*, \gamma^*) \nabla \mathbf{m}_g(\theta^0) \right] \\ & + \frac{1}{G} \sum_g \sum_{h, h \neq g} \mathbb{E} \left[ \nabla \mathbf{m}_g^\top(\theta^0) \mathbf{W}_g^{-1}(\theta^*, \gamma^*) \mathbf{u}_g \mathbf{u}_h^\top \mathbf{W}_h^{-1}(\theta^*, \gamma^*) \nabla \mathbf{m}_h(\theta^0) \right]. \end{aligned} \quad (50)$$

and  $\mathbf{AS}_\infty = \lim_{G \rightarrow \infty} AS_G$ . We then impose:

**Assumption A.8.** (Variance-covariance matrix). *There exist two positive constants,  $c'$  and  $C'$ , such that  $c' < \lambda_{\min}(\mathbb{E}[\nabla \mathbf{m}_g^\top(\theta^0) \mathbf{W}_g^{-1}(\theta^*, \gamma^*) \nabla \mathbf{m}_g(\theta^0)]) < \lambda_{\max}(\mathbb{E}[\nabla \mathbf{m}_g^\top(\theta^0) \mathbf{W}_g^{-1}(\theta^*, \gamma^*) \nabla \mathbf{m}_g(\theta^0)]) < C'$ . Furthermore,  $\inf_G |D_G|^{-1} \lambda_{\min}(\mathbf{AS}_\infty) > 0$ .*

**Assumption A.9.** (Initial estimator and root assumption). *It holds that  $\mathbf{S}_G(\check{\gamma}, \check{\theta}, \hat{\theta}) = \mathcal{O}_p(1/\sqrt{G})$ .*

Assumption A.6 is needed for establishing moment inequalities which are necessary for establishing CLT; see also Assumption 10 in Xu and Lee (2015a). Assumption A.7 concerns the pre-estimation of the nuisance parameters  $\gamma$  and  $\theta$ ; It shall be noted that in Assumption A.7,  $(\theta^*, \gamma^*)$  can be different from the

true parameter  $(\theta^0, \gamma^0)$ , which will not affect the consistency of our estimator. For more explanations, see the proof of [Theorem 2](#) in supplementary material. In Section S1.4 of supplementary material, we also provide primitive conditions for verifying that the estimator for  $\check{\gamma}$ , presented by (12) in [Section 2](#), satisfies [Assumption A.7](#). [Assumption A.8](#) is a standard regularity condition for nonlinear estimation. The assumption on score function in [Assumption A.9](#) acknowledges the fact that for some nonlinear estimation equations, the existence of a solution might not be a trivial issue; see [Jacod and Sørensen \(2018\)](#) for relevant discussions.

We define  $\mathbf{H}_\infty = \lim_{G \rightarrow \infty} \mathbb{E} \mathbf{H}_G(\theta^*, \gamma^*, \theta^0)$ , where  $\mathbf{H}_G(\theta^*, \gamma^*, \theta^0)$  is the Hessian matrix with respect to  $\theta_0$ . Furthermore, define  $\mathbf{AV} = \mathbf{AV}(\gamma^*, \theta^*, \theta^0) \stackrel{\text{def}}{=} \mathbf{H}_\infty^\top \mathbf{AS}_\infty \mathbf{H}_\infty$ . It is not surprising to see in the following [Theorem 2](#) that our estimation is asymptotically normal, and the rate of convergence is  $\sqrt{G}$ .

**Theorem 2.** *Under Assumptions A.1–A.9, we have*

$$\sqrt{G} \mathbf{AV}^{-1/2} (\hat{\theta} - \theta^0) \sim_d \mathbb{N}(0, I_p). \quad (51)$$

#### 4.5. Consistency of variance-covariance matrix estimation

In this subsection, we propose a semiparametric estimator of the asymptotic variance in [Theorem 2](#) and prove its consistency. The estimation is tailored to account for the spatial dependence of the underlying process. This facilitates the statistical inference for the proposed estimator.

First, we define

$$\hat{\mathbf{A}} = \frac{1}{|D_G|} \sum_g \nabla \hat{\mathbf{m}}_g^\top \hat{\mathbf{W}}_g^{-1} \nabla \hat{\mathbf{m}}_g, \quad (52)$$

$$\hat{\mathbf{B}} = \frac{1}{|D_G|} \sum_g \sum_{h \neq g} k(d_{gh}) \nabla \hat{\mathbf{m}}_g^\top \hat{\mathbf{W}}_g^{-1} \hat{\mathbf{u}}_g \hat{\mathbf{u}}_h^\top \hat{\mathbf{W}}_h^{-1} \nabla \hat{\mathbf{m}}_h^\top, \quad (53)$$

where  $\nabla \hat{\mathbf{m}}_g \equiv \nabla \hat{\mathbf{m}}_g(\hat{\theta})$ ,  $\hat{\mathbf{W}}_g \equiv \hat{\mathbf{W}}_g(\check{\gamma}, \check{\theta})$ , and  $\hat{\mathbf{u}}_g = \mathbf{y}_g - \mathbf{m}_g(\hat{\theta})$  is an estimator for  $\mathbf{u}_g$ .

Following the spatial HAC literature (see, e.g., [Kelejian and Prucha, 2007](#)), we construct an estimator for  $\mathbf{AV}$  as

$$\begin{aligned} \widehat{\mathbf{AV}}(\check{\gamma}, \check{\theta}, \hat{\theta}) &= |D_G| \left( \sum_g \nabla \hat{\mathbf{m}}_g^\top \hat{\mathbf{W}}_g^{-1} \nabla \hat{\mathbf{m}}_g \right)^{-1} \\ &\quad \left( \sum_g \sum_{h(\neq g)} \nabla \hat{\mathbf{m}}_g^\top \hat{\mathbf{W}}_g^{-1} k(d_{gh}) \hat{\mathbf{u}}_g \hat{\mathbf{u}}_h^\top \hat{\mathbf{W}}_h^{-1} \nabla \hat{\mathbf{m}}_h \right) \\ &\quad \left( \sum_g \nabla \hat{\mathbf{m}}_g^\top \hat{\mathbf{W}}_g^{-1} \nabla \hat{\mathbf{m}}_g \right)^{-1} \\ &= \hat{\mathbf{A}}^{-1} \hat{\mathbf{B}} \hat{\mathbf{A}}^{-1}, \end{aligned} \quad (54)$$

where  $k(d_{gh})$  is a kernel function depending on the distance between group  $g$  and  $h$ , i.e.,  $d_{gh}$  (which is also denoted as  $d(g, h)$ ), and a bandwidth parameter  $b_g$ . As noted in [Kelejian and Prucha \(2007\)](#), there are many choices for the kernel function, such as the rectangular kernel, Bartlett or triangular kernel. In particular, without loss of generality, we can choose the Bartlett kernel function:  $k(d_{gh}) = 1 - d(g, h)/b_g$  for  $d(g, h) < b_g$ , and  $k(d_{gh}) = 0$  for  $d(g, h) \geq b_g$ .

We now list assumptions that are needed for the consistent estimation of the variance-covariance matrix.

**Assumption A.10.** (Residual property).  $\hat{\mathbf{u}}_g - \mathbf{u}_g = C_g \Delta_g$ , where  $C_g$  is a  $L \times p$  matrix, and  $\Delta_g$  is a  $p \times 1$  dimensional vector. It holds that  $\|C_g\|_2 = \mathcal{O}_p(1)$  and  $\|\Delta_g\| = \mathcal{O}_p(G^{-1/2})$ .

**Assumption A.11.** (Kernel assumption). The kernel function  $k(\cdot)$  satisfies  $|k(d_{gh}) - 1| \leq C_k |d_{gh}/b_g|^{\rho_k}$ , where  $\rho_k > 0$ ,  $d_{gh}/b_g \leq 1$ ,  $b_g \rightarrow 0$ , and  $C_k$  is a generic constant. It also holds that  $b_g^{d/q'} |D_G|^{-1} = \mathcal{O}(1)$ ,  $|D_G|^{-1} \sum_g \sum_h |d(g, h)/b_g|^{\rho_k} = \mathcal{O}(1)$ , and  $b_g^{2d} \sum_{r=1}^{\infty} r^{d-1} \psi((r - b_g)_+) = \mathcal{O}(G)$ , where  $(r - b_g)_+ = \max(r - b_g, 0)$ , and recall that  $\psi(\cdot)$  is the NED coefficient.

Assumption B.1 is an assumption for decomposing the difference between the residuals and the true error; a similar assumption is imposed by Kelejian and Prucha (2007). Assumption B.2 concerns the properties of the kernel function, and also puts constraints on the spatial dependence coefficients and the bandwidth parameter. In the following theorem, we demonstrate the consistency of the  $\widehat{AV}$ .

**Theorem 3.** Under Assumption A.1–A.9 and B.1–B.2, the variance-covariance estimator presented in (54) is consistent, i.e.,  $\widehat{AV}(\check{\gamma}, \check{\theta}, \hat{\theta}) \xrightarrow{P} AV(\gamma^*, \theta^*, \theta^0)$ .

## 5. Monte Carlo simulations

In this section, we use Monte Carlo simulations to investigate the finite sample performance of the proposed GEE approach and compare it to QMLE. By studying five different cases for count data and binary response data, we show that the proposed GEE estimator not only has better performance under various data-generating processes, but also is robust to moderate misspecification of the working correlation matrix or the group structure. Our code has been uploaded to GitHub (link: <https://github.com/Uwe-xu/NonlinearGEE>.)

### 5.1. Sampling space

We sample 400 or 1600 observations on a linear lattice. The data are divided into groups of size 4, with each group's points being normally distributed, sharing the same mean and a variance of 0.1. In the case where  $n = 400$ , the group means are represented by 100 equally spaced points in  $[0, 10]$ , while in the case where  $n = 1600$ , they are represented by 400 equally spaced points. The distance  $d_{ij}$  between locations  $i$  and  $j$  is calculated as Euclidean distance on the real line.

### 5.2. Count data

#### 5.2.1. Data-generating process

For a spatial Poisson distribution, given the spatial correlation, the variances and covariances of the count-dependent variable can be expressed in closed forms, as illustrated in Eqs. (22) and (23). Consider the following spatial count data-generating process:

$$\begin{aligned} m_i &= \xi_i \exp(\beta_1 x_{i,1} + \beta_2 x_{i,2}), \\ \beta_1 &= \beta_2 = 1, \\ x_{i,1}, x_{i,2} &\stackrel{\text{i.i.d.}}{\sim} \text{N}(0, 1), \\ Y_i &\sim \text{Poisson}(m_i), \end{aligned}$$

where  $\xi_i$  is a random variable independent of  $X$  with  $E(\xi_i) = 1$ . We consider two cases with different types of spatial correlations of  $\xi_i$ . In the following, we describe the case-specific data-generating processes and within-group correlation matrices. The corresponding variance-covariance matrix,  $\mathbf{W}_g$ , is defined in Eq. (9), and estimation procedures can be found in Section 3.

**Case 1.**  $\xi_i$  is simulated as a multivariate lognormal variable by exponentiating an underlying multivariate normal distribution that has a marginal distribution  $N(-\frac{1}{2}, 1)$  and a within-group correlation matrix  $\Omega_g$ . For  $i \neq j$ , we have  $\Omega_{g,ij} = \rho$ , i.e., the correlation matrix is exchangeable. We let  $\rho = 0.1, 0.5, 0.8$ , and  $1$ . The underlying normal distribution implies that  $\xi_i$  follows a multivariate lognormal distribution with  $E(\xi_i) = 1$ . The group size equals four. The within-group correlation matrix is

$$\Omega_g = \begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix}. \quad (55)$$

**Case 2.** The setting is similar to Case 1: we have  $\xi_i \sim \log N(-\frac{1}{2}, 1)$ , and the group size is four. However, the correlation between  $j$ -th and  $i$ -th elements is specified as  $\text{corr}(\xi_i, \xi_j) = \rho(1 - d_{ij})$ , i.e., it depends on both  $\rho$  and the spatial distance. The within-group correlation matrix is

$$\Omega_g = \begin{pmatrix} 1 & \rho(1 - d_{12}) & \rho(1 - d_{13}) & \rho(1 - d_{14}) \\ \rho(1 - d_{21}) & 1 & \rho(1 - d_{23}) & \rho(1 - d_{24}) \\ \rho(1 - d_{31}) & \rho(1 - d_{32}) & 1 & \rho(1 - d_{34}) \\ \rho(1 - d_{41}) & \rho(1 - d_{42}) & \rho(1 - d_{43}) & 1 \end{pmatrix}. \quad (56)$$

## 5.2.2. Simulation results

Table 1 shows simulation results under Case 1 and Case 2, with two sample sizes:  $n \in \{400, 1600\}$  and  $G \in \{100, 400\}$ . There are two estimators, Poisson QMLE and Poisson GEE. For the Poisson GEE estimation, we divide the data into groups of four, sequentially grouping every four elements from start to finish. Then, we employ the estimated working variance-covariance matrices (55) and (56) in the GEE estimation for Cases 1 and 2, respectively. All the columns are based on 1000 Monte Carlo replications. For each setting, Monte Carlo means and standard errors are computed.

In general, we can see that the Monte Carlo means are very close to the true values, suggesting that in Case 1 and Case 2, both methods are asymptotically unbiased. Therefore, we can focus on the comparison of standard errors. We use bold font to highlight estimates that have smaller standard errors.

From Table 1, we see that the Poisson GEE methods perform considerably better than Poisson QMLE methods in every case, sample size, and  $\rho$ . For example, in Case 1 and  $n = 400$ , the improvements (in terms of the standard deviation) range from 28% (for  $\hat{\beta}_2$  with  $\rho = 0.1$ ) to 50% (for  $\hat{\beta}_2$  with  $\rho = 1$ ); in Case 2 and  $n = 1600$ , the improvements range from 37% (for  $\hat{\beta}_1$  with  $\rho = 0.1$ ) to 66% (for  $\hat{\beta}_1$  with  $\rho = 1$ ).

Another observation is that with  $\rho$  increasing, the performance gaps between the two methods become larger. For example, in Case 1 and  $n = 400$ , when  $\rho = 0.1$  (weak spatial correlation), the improvement for  $\hat{\beta}_1$  is  $1 - 0.1061/0.1486 \approx 29\%$ ; when  $\rho = 1$  (high spatial correlation), the improvement for  $\hat{\beta}_1$  increases to  $1 - 0.08/0.1489 \approx 44\%$ . This is in line with our expectation since the proposed GEE method is capable of accounting for the spatial correlation in the underlying innovations. The stronger the correlation, the larger the improvement.

## 5.3. Binary response data

### 5.3.1. Data-generating process and misspecified working correlation matrix

For the spatial Probit model, the correlations of the latent errors result in the correlations of the observable binary response variables. However, the correlations become different after transformation, and the true correlation between two binary responses does not have a closed analytical form. This provides a natural setting to test the robustness of the proposed GEE methods when the working correlation matrix is moderately misspecified. Let us consider the following setting:

**Table 1.** Means and standard deviations for Case 1 and Case 2, averaged over 1000 samples.

		$n = 400, G = 100, L = 4$				$n = 1600, G = 400, L = 4$			
		Case 1		Case 2		Case 1		Case 2	
		Poisson	GEE-Poisson	Poisson	GEE-Poisson	Poisson	GEE-Poisson	Poisson	GEE-Poisson
$\rho = 1$	$\hat{\beta}_1$	1.0475	1.0067	1.0593	0.9994	0.9993	0.9956	1.0012	0.9959
	s.d. ( $\hat{\beta}_1$ )	0.1489	<b>0.0800</b>	0.1494	<b>0.0666</b>	0.0804	<b>0.0295</b>	0.0805	<b>0.0275</b>
	$\hat{\beta}_2$	1.0465	1.0078	1.0557	1.0015	1.0007	0.9966	1.0037	0.9963
	s.d. ( $\hat{\beta}_2$ )	0.1494	<b>0.0750</b>	0.1470	<b>0.0696</b>	0.0790	<b>0.0295</b>	0.0808	<b>0.0283</b>
$\rho = 0.8$	$\hat{\beta}_1$	1.0408	0.9900	1.0620	0.9966	1.0019	0.9963	0.9827	0.9882
	s.d. ( $\hat{\beta}_1$ )	0.1511	<b>0.0955</b>	0.1558	<b>0.1025</b>	0.0810	<b>0.0414</b>	0.0752	<b>0.0386</b>
	$\hat{\beta}_2$	1.0448	0.9886	1.0580	0.9925	1.0021	0.9923	0.9855	0.9904
	s.d. ( $\hat{\beta}_2$ )	0.1596	<b>0.0988</b>	0.1536	<b>0.0979</b>	0.0759	<b>0.0424</b>	0.0739	<b>0.0356</b>
$\rho = 0.5$	$\hat{\beta}_1$	1.0325	0.9893	1.0431	1.0062	0.9970	0.9968	1.0162	1.0058
	s.d. ( $\hat{\beta}_1$ )	0.1509	<b>0.1120</b>	0.1397	<b>0.1111</b>	0.0752	<b>0.0429</b>	0.0747	<b>0.0406</b>
	$\hat{\beta}_2$	1.0375	0.9939	1.0344	1.0032	1.0012	0.9979	1.0172	1.0086
	s.d. ( $\hat{\beta}_2$ )	0.1454	<b>0.1123</b>	0.1433	<b>0.1067</b>	0.0744	<b>0.0424</b>	0.0736	<b>0.0447</b>
$\rho = 0.1$	$\hat{\beta}_1$	1.0225	1.0046	0.9567	0.9556	1.0013	0.9996	1.0007	0.9967
	s.d. ( $\hat{\beta}_1$ )	0.1486	<b>0.1061</b>	0.1470	<b>0.1125</b>	0.0750	<b>0.0453</b>	0.0791	<b>0.0468</b>
	$\hat{\beta}_2$	1.0199	1.0036	0.9554	0.9494	0.9981	0.9968	1.0018	0.9999
	s.d. ( $\hat{\beta}_2$ )	0.1439	<b>0.1039</b>	0.1455	<b>0.1102</b>	0.0800	<b>0.0462</b>	0.0761	<b>0.0481</b>

s.d. stands for standard deviations.  
 The estimates with smaller standard deviations are highlighted in bold font.

**Case 3**

$$y_i = 1(y_i^* > 0), \tag{57}$$

$$y_i^* = \mathbf{x}_i\beta + e_i, \tag{58}$$

$$u_i = y_i - E(y_i|x_i) = y_i - \Phi(\mathbf{x}_i\beta),$$

where the true parameter  $\beta = (\beta_1, \beta_2)^T = (1, 1)^T$  and  $\mathbf{x}_i \sim N(0, I_2)$ ; The marginal distribution of  $e_i$  is  $N(0, 1)$ . With the group size  $L = 4$ , the correlation between  $i$ -th and  $j$ -th elements is  $corr(e_i, e_j) = \rho(1 - d_{ij})$  if they are in the same group, and 0 otherwise.

We specify the following conditional working correlation matrix for  $u_i$  in group  $g$ :

$$\Omega_g = \begin{pmatrix} 1 & c(1 - d_{12}) & c(1 - d_{13}) & c(1 - d_{14}) \\ c(1 - d_{21}) & 1 & c(1 - d_{23}) & c(1 - d_{24}) \\ c(1 - d_{31}) & c(1 - d_{32}) & 1 & c(1 - d_{34}) \\ c(1 - d_{41}) & c(1 - d_{42}) & c(1 - d_{43}) & 1 \end{pmatrix}. \tag{59}$$

While  $\Omega_g$  might mimic the correlation structure of  $e_i$ , it does not correctly specify that of  $u_i$ . Namely, the working correlation matrix is misspecified. The parameter  $c$  can be regarded as the coefficient of the linear regression of the product of the group member  $u_i$  and  $u_j$  on their closeness  $(1 - d_{ij})$ . In estimation, we replace  $u_i$  with the plug-in  $\check{u}_i := y_i - \Phi(\mathbf{x}_i\check{\beta}_{\text{QMLE}})$ .

**5.3.2. Simulation results**

Table 2 shows the simulation results of Case 3 with two sample sizes and group numbers:  $n \in \{400, 1600\}$  and  $G \in \{100, 400\}$ . The two estimators are Probit QMLE and Probit GEE. For the Probit GEE approach, the grouping strategy is implemented in the same way as described in Section 5.2.2. All the columns are based on 1000 Monte Carlo replications. In general, we can see that for Case 3, the estimation biases are in general larger than those in Case 1 and Case 2. Thus, we choose mean square errors (MSE) instead of standard deviations to measure the estimation performance. We highlight the estimates with smaller MSE in bold font.

When the spatial correlation is comparatively large ( $\rho = 1, 0.8$ , and  $0.5$  in the latent model), the Probit GEE performs better than Probit QMLE in every sample size. For example, for  $n = 400$ , the

**Table 2.** Means and MSEs for Case 3 averaged over 1000 samples

		$n = 400, G = 100, L = 4$		$n = 1600, G = 400, L = 4$	
		Probit	GEE-Probit	Probit	GEE-Probit
$\rho = 1$	$\hat{\beta}_1$	0.8880	0.8897	1.0046	1.0055
	MSE( $\hat{\beta}_1$ )	8.1876	<b>7.4890</b>	4.1530	<b>3.3511</b>
	$\hat{\beta}_2$	0.8858	0.8874	1.0061	1.0067
$\rho = 0.8$	MSE( $\hat{\beta}_2$ )	8.5451	<b>7.8048</b>	4.4365	<b>3.5446</b>
	$\hat{\beta}_1$	0.9087	0.9099	0.9938	0.9950
	MSE( $\hat{\beta}_1$ )	6.7392	<b>6.1236</b>	3.9748	<b>3.5374</b>
$\rho = 0.5$	$\hat{\beta}_2$	0.9047	0.9055	0.9960	0.9968
	MSE( $\hat{\beta}_2$ )	7.1594	<b>6.6605</b>	4.0776	<b>3.5528</b>
	$\hat{\beta}_1$	0.9372	0.9380	0.9831	0.9831
$\rho = 0.1$	MSE( $\hat{\beta}_1$ )	5.3103	<b>5.0167</b>	4.5959	<b>4.3844</b>
	$\hat{\beta}_2$	0.9368	0.9372	0.9852	0.9853
	MSE( $\hat{\beta}_2$ )	5.2776	<b>5.0261</b>	4.1201	<b>3.8726</b>
$\rho = 0.1$	$\hat{\beta}_1$	0.9771	0.9774	0.9769	0.9776
	MSE( $\hat{\beta}_1$ )	<b>4.4622</b>	4.5111	4.7612	<b>4.5880</b>
	$\hat{\beta}_2$	0.9671	0.9671	0.9768	0.9775
	MSE( $\hat{\beta}_2$ )	<b>4.2051</b>	4.3404	4.7526	<b>4.5802</b>

MSE stands for mean square errors.  
The estimates with smaller MSE are highlighted in bold font.

improvements (in terms of MSE) range from 4.8% (for  $\beta_2$  with 0.5) to 8.7% (for  $\beta_2$  with  $\rho = 1$ ); for  $n = 1600$ , the improvements range from 4.6% (for  $\beta_2$  with  $\rho = 0.5$ ) to 19.3% (for  $\beta_1$  with  $\rho = 1$ ). Again, the improvement increases with the (latent) spatial correlation parameter  $\rho$ . When the spatial correlation is very small and the sample size is relatively small ( $\rho = 0.1$  in the latent model and  $n = 400$ ), Probit QMLE has smaller MSEs than Probit GEE.

In sum, we can conclude that the Probit GEE outperforms Probit QMLE in most cases. We also note that in general, the improvements are smaller than those in count models. There are two reasons: (i) We use a moderately misspecified working correlation matrix (59), which only captures a part of the spatial correlations in errors  $u_i$ ; (ii) the transformation from the latent model (58) to the observed model (57) attenuates the correlation. This explains the unexpected result for the case of  $\rho = 0.1$  and  $n = 400$ : The original spatial correlation in the latent model is weak, resulting in a weaker correlation in the observed model. Furthermore, the small sample size makes the situation worse. If we increase the sample size to 1600, Probit GEE becomes the best of the two methods again.

### 5.4. Misspecified group

In practice, the group structure can be misspecified, particularly if the data does not naturally segregate into distinct groups. In this section, we will concentrate on the cases where the groups defined in the estimation process do not align with the group structure inherent to the data-generating processes.

In Case 4 and Case 5, we follow the model in Case 3

$$\begin{aligned}
 y_i &= 1 (y_i^* > 0) \\
 y_i^* &= \mathbf{x}_i \beta_0 + e_i \\
 u_i &= y_i - \Phi(\mathbf{x}_i \beta_0),
 \end{aligned}$$

but the group sizes for the data-generating processes are  $L = 2$  for Case 4 and  $L = 8$  for Case 5. During the estimation, we continue to employ the same specifications of the working correlation matrix as previously done in Case 3, implementing a group size of  $L = 4$  (cf. Section 5.2.2). Under this estimation strategy, the groups are misspecified in both cases. In Case 4, the groups specified in the estimation are

**Table 3.** Means and standard deviations for the binary cases 4 and 5, averaged over 1000 samples

		$n = 400$				$n = 1600$			
		$L = 2, G = 200$		$L = 8, G = 50$		$L = 2, G = 800$		$L = 8, G = 200$	
		Case 4		Case 5		Case 4		Case 5	
		Probit	GEE-Probit	Probit	GEE-Probit	Probit	GEE-Probit	Probit	GEE-Probit
$\rho = 1$	$\hat{\beta}_1$	0.9034	0.9025	0.8964	0.8967	0.9954	0.9954	1.0277	1.0296
	MSE( $\hat{\beta}_1$ )	7.4279	<b>7.2388</b>	7.6083	<b>6.8683</b>	4.4444	<b>4.0655</b>	5.6752	<b>4.8035</b>
	$\hat{\beta}_2$	0.9050	0.9050	0.9007	0.8984	0.9926	0.9928	1.0280	1.0301
	MSE( $\hat{\beta}_2$ )	7.2592	<b>6.9608</b>	7.1524	<b>6.7493</b>	<b>4.0545</b>	4.0676	5.3047	<b>4.7150</b>
$\rho = 0.8$	$\hat{\beta}_1$	0.9151	0.9146	0.9107	0.9115	0.9919	0.9920	1.0118	1.0127
	MSE( $\hat{\beta}_1$ )	6.6283	<b>6.5599</b>	6.5958	<b>6.0885</b>	4.3971	<b>4.1647</b>	4.8029	<b>4.0730</b>
	$\hat{\beta}_2$	0.9164	0.9169	0.9117	0.9117	0.9882	0.9880	1.0138	1.0152
	s.d.( $\hat{\beta}_2$ )	<b>6.2610</b>	6.2941	6.6896	<b>6.2341</b>	4.1810	<b>4.0104</b>	4.4290	<b>3.8600</b>
$\rho = 0.5$	$\hat{\beta}_1$	0.9267	0.9270	0.9223	0.9228	0.9877	0.9879	0.9953	0.9956
	MSE( $\hat{\beta}_1$ )	5.7652	<b>5.6154</b>	5.9215	<b>5.5827</b>	4.6214	<b>4.5277</b>	4.3116	<b>3.9813</b>
	$\hat{\beta}_2$	0.9311	0.9314	0.9226	0.9228	0.9844	0.9844	0.9968	0.9976
	MSE( $\hat{\beta}_2$ )	5.3039	<b>5.2009</b>	5.7635	<b>5.4962</b>	4.5617	<b>4.5057</b>	4.0585	<b>3.7314</b>
$\rho = 0.1$	$\hat{\beta}_1$	0.9521	0.9521	0.9540	0.9545	0.9822	0.9822	0.9776	0.9775
	MSE( $\hat{\beta}_1$ )	4.9518	<b>4.9357</b>	<b>4.2188</b>	4.2327	<b>4.5033</b>	4.5048	<b>4.8236</b>	4.8293
	$\hat{\beta}_2$	0.9539	0.9541	0.9533	0.9530	0.9797	0.9796	0.9778	0.9778
	MSE( $\hat{\beta}_2$ )	<b>4.8634</b>	4.8948	<b>4.7074</b>	4.7704	4.8286	<b>4.8186</b>	<b>4.5974</b>	4.6024

MSE stands for the mean square error.

The estimates with smaller MSEs are highlighted in bold font.

The group structure implemented in estimation:  $L = 4$  and  $G = n/L$ .

too coarse, leading to pairs of observations, which actually belong to different groups, being clustered into the same group. As a result, the proposed GEE method should become less efficient since it estimates and plugs in many estimators of correlations that are zero, thereby introducing additional estimation noises. In Case 5, the groups specified in the estimation are overly granular, causing many pairs of observations, which inherently belong to the same group, to be categorized into different groups. Consequently, their correlations are misspecified as 0.

Moreover, we want to emphasize that our estimation strategy in both Case 4 and Case 5 confronts misspecification issues in *both* within-group structure and the group size. The former arises from the latent model transformation, as discussed in Case 3 (following Eq. (59)).

#### 5.4.1. Simulation results

Table 3 presents the simulation results. In general, the proposed GEE method outperforms the Probit-QMLE when the groups are moderately misspecified. We implemented 1,000 Monte Carlo simulations across eight settings, estimating a total of 16 parameters (2 cases  $\times$  4  $\rho$ 's  $\times$  2 methods). In Case 4, of these 16 parameters, the Probit-GEE achieves smaller Monte-Carlo MSEs 12 times. In contrast, the Probit-QMLE yields smaller MSEs only three times, two of which occur in the case of  $\rho = 0.1$ . In this instance, the spatial correlation is minimal, rendering the Probit-QMLE nearly optimal.

The results for Case 5 follow a similar pattern. Out of the estimation of 16 parameters, our proposed method attains smaller Monte-Carlo MSEs 12 times, and the majority of scenarios where the Probit-QMLE outperforms the Probit-GEE occur in the case of  $\rho = 0.1$ .

Overall, the simulation results support the implementation of the proposed GEE method.

## 6. An empirical application of the role of cultural distance in the gravity equation

The gravity equation has been widely used in international trade since Tinbergen (1962). In this section, we extend the gravity equation by incorporating cultural distance (CD) between countries and demonstrate how to use the proposed GEE method to estimate the extended gravity equation.



## 6.1. Background and notations

Anderson and Van Wincoop (2003) specify the gravity equation as:

$$T_{ij} = \alpha_0 Y_i^{\alpha_1} Y_j^{\alpha_2} D_{ij}^{\alpha_3} \eta_{ij}, \quad (60)$$

where  $T_{ij}$  is the trade flow between country  $i$  and country  $j$ ;  $T_{ij}$  is proportional to the product of the two countries' GDPs, denoted by  $Y_i$  and  $Y_j$ , and inversely proportional to their distance.  $D_{ij}$  broadly represents trade resistance, while  $\eta_{ij}$  is a multiplicative stochastic error. In the literature, it is traditional to take the natural logarithms of both sides of an equation and to include additional control variables, represented by  $Z_{ij}$ . The resulting log-linearized equation is as follows:

$$\log T_{ij} = \log \alpha_0 + \alpha_1 \log Y_i + \alpha_2 \log Y_j + \alpha_3 \log D_{ij} + \beta Z_{ij} + \log \eta_{ij}. \quad (61)$$

A traditional estimation approach for (61) is to use ordinary least squares (OLS). However, log-linearized models estimated by OLS can be highly misleading in the presence of heteroscedasticity. Silva and Tenreyro (2006) discuss this situation and suggest using a nonlinear estimator which is numerically equivalent to the pseudo Poisson MLE. Their approach is essentially a pooled Poisson QMLE that does not account for any spatial correlation. In this section, we adopt nonlinear specification to the gravity equation and further apply the proposed GEE approach using the product-level trade data between China and the rest of the world in 2016. By including the culture distance variable ( $\log CD_{gl}$ ), we specify the conditional mean function of the trade volume as follows:

$$\begin{aligned} E(\text{Trade}_{glk} | X_{gl}) = & \exp(\beta_0 + \beta_1 \log CD_{gl} + \beta_2 \log GDPPC_{gl} + \beta_3 \log Dist_{gl} \\ & + \beta_4 \text{Language}_{gl} + \beta_5 \text{Landlock}_{gl}), \end{aligned} \quad (62)$$

where  $g = 1, 2, \dots, G$  index groups, and  $l = 1, 2, \dots, L_g$  index the members in the group  $g$ , and  $k$  is an index for products.  $\text{Trade}_{gl}$  is the trade flow between China and the  $l$ th country in group  $g$ . We study three types of the dependent variable of  $\text{Trade}_{gl}$ : export, import, and total trade. The control variables include the log of country  $gl$ 's GDP per capita ( $\log GDPPC_{gl}$ ), the geographical distance from China to country  $gl$  ( $\log Dist_{gl}$ ), an indicator as to whether country  $gl$  shares a common language with China ( $\text{Language}_{gl}$ ), and an indicator as to whether country  $gl$  is landlocked ( $\text{Landlock}_{gl}$ ).

## 6.2. Data source

The data on trade volume are from Trade Map ([www.trademap.org](http://www.trademap.org)). Per capital GDP is from World Development Indicators (WDI). Geographical distance, common language, and landlock are from the French Centre d'Etudes Prospectives et d'Informations Internationales (CEPII). Cultural distance is calculated based on the dataset of Hofstede Insights, which provides national culture scores for 103 countries or regions in the year 2016. Due to missing data, our final sample contains pairwise trade data between China and 95 other countries across 97 products, resulting in a sample size of 9215 observations.

In particular, the cultural distance measure,  $CD_{gl}$ , is calculated using six different national culture scores: power distance, individualism versus collectivism, masculinity versus femininity, uncertainty avoidance, long-term versus short-term normative orientation, and indulgence versus restraint. We refer to Hofstede, Hofstede, and Minkov (2010) for more details about the calculation of cultural distance.

## 6.3. Estimation strategies

In Section 6.4.1 below, we apply OLS, Poisson QMLE, and Poisson GEE to the aforementioned dataset and compare their performance. The Poisson QMLE is implemented as suggested by Silva and Tenreyro (2006). In the OLS approach, we estimate a model that is the log-linearized version of (62):

$$\begin{aligned} \log \text{Trade}_{glk} = & \beta_0 + \beta_1 \log(CD_{gl}) + \beta_2 \log(GDPPC_{gl}) + \beta_3 \log(Dist_{gl}) \\ & + \beta_4 \text{Language}_{gl} + \beta_5 \text{Landlock}_{gl} + u_{gl}. \end{aligned}$$

**Table 4.** OLS estimates of the gravity equation.

	Trade	Export	Import
CD	−0.432***	−0.345***	−0.557***
s.e.	(0.043)	(0.043)	(0.064)
logGDPPC	0.547***	0.398***	0.989***
s.e.	(0.028)	(0.029)	(0.045)
logDist	−0.912***	−0.993***	−1.073***
s.e.	(0.069)	(0.069)	(0.097)
Language	0.109	0.439*	−1.305***
s.e.	(0.177)	(0.179)	(0.269)
Landlock	−2.130***	−2.213***	−1.095***
s.e.	(0.110)	(0.111)	(0.151)
constant	13.645***	15.090***	8.868***
s.e.	(0.680)	(0.683)	(0.942)
N	8513	8318	5737

s.e. stands for heteroskedasticity robust standard errors.

For the proposed GEE approach, a group structure has to be chosen for constructing the working variance-covariance matrix. Given the data is inherently divided into 97 groups by product, we use these product types to define the groups in the following [Section 6.4.1](#). Without further information to differentiate products within each group, a natural choice for the within-group dependence structure is the exchangeable correlation matrix; see [Eq. \(55\)](#) for an example.

Furthermore, as a robustness check, we apply the proposed GEE approach to country-level data in the subsequent [Section 6.4.2](#). The country-level data aggregates the trade volumes of 97 products between a country and China into a single item, resulting in a sample of 95 countries. (As a result, we remove the group index  $k$  from the model [\(62\)](#).) In this separate analysis, we group the data based on geographical locations. In a baseline setup, we divide countries into different groups according to their continents. This natural strategy results in five major groups (w.r.t. five continents). In the second grouping, we divide countries into groups, each containing six members. Specifically, we group every six countries from the same continent together in the initial round. Then, if there are remaining countries within a continent, we incorporate these leftovers with those from other continents into the last group. Given the sample of 95 countries, the second strategy results in 16 groups. To test the robustness of our approach under different (possibly misspecified) within-group dependence structures, we continue to use the exchangeable correlation matrix. It is worth noting that the proposed GEE approach can deliver decent performance even when the working variance-covariance matrix is moderately misspecified; see [Section 2.4](#) and [Sections 5.3](#) and [5.4](#) for theoretical discussions and simulation evidence, respectively.

## 6.4. Estimation results

### 6.4.1. Main results at the product level

For comparison, we provide the OLS estimates of the log-linearized model in [Table 4](#).

The log-linearized model suffers from two main problems. First, the dependent variable cannot be log-transformed if it is zero. As a result, the sample size in each column in [Table 4](#) has reduced a lot, especially in the import column. Second, as mentioned in [Silva and Tenreiro \(2006\)](#), the log-linearization can cause bias in parameter estimates if there is heteroskedasticity in the error term. In comparison, the Poisson QMLE and Poisson GEE are less likely to be prone to these biases. The pooled Poisson QMLE estimates, which do not account for any spatial correlation, are provided in [Table 5](#); the estimates from the proposed GEE approach are provided in [Table 6](#).

Comparing [Tables 5](#) and [6](#), we see that the Poisson GEE method generally reduces the standard errors of parameter estimates. All estimation results confirm the negative effect of the cultural distance on

**Table 5.** Poisson QMLE estimates of the gravity equation.

	Trade	Export	Import
logCD	−0.378***	−0.306***	−0.493**
s.e.	(0.083)	(0.059)	(0.164)
logGDPPC	0.760***	0.699***	0.860***
s.e.	(0.093)	(0.094)	(0.148)
logDist	−0.665***	−0.609***	−0.729***
s.e.	(0.157)	(0.145)	(0.211)
Language	−0.522	0.030	−1.611*
s.e.	(0.445)	(0.428)	(0.711)
Landlock	−1.832***	−2.354***	−1.439*
s.e.	(0.401)	(0.274)	(0.567)
constant	12.359***	11.769***	11.332***
s.e.	(1.781)	(1.890)	(2.357)
N	9215	9215	9215

s.e. stands for heteroskedasticity robust standard errors.

**Table 6.** Poisson GEE estimates of the gravity equation.

	Trade	Export	Import
logCD	−0.381***	−0.327***	−0.515***
s.e.	(0.054)	(0.012)	(0.130)
logGDPPC	0.766***	0.742***	0.895***
s.e.	(0.077)	(0.069)	(0.183)
logDist	−0.667***	−0.627***	−0.738**
s.e.	(0.143)	(0.093)	(0.255)
Language	−0.529	−0.004	−1.672**
s.e.	(0.303)	(0.211)	(0.526)
Landlock	−1.848***	−2.548***	−1.500*
s.e.	(0.399)	(0.189)	(0.613)
constant	12.317***	11.450***	11.066***
s.e.	(1.802)	(1.052)	(3.352)
N	9215	9215	9215

s.e. stands for standard errors that are robust to model misspecification.

trade in the gravity equation, along with the negative effect of the geographical distance and the positive effect of the GDP per capita. These estimates are all significant at the 1% level. For the main explanatory variable, the cultural distance, the Poisson GEE estimated coefficients for trade, export, and import are −0.381, −0.327, and −0.515. All of them are larger in absolute values than the corresponding effects estimated with Poisson QMLE. Furthermore, all the standard errors of the Poisson GEE estimates for the coefficients of the cultural distance are smaller than those of Poisson QMLE. Overall, the Poisson GEEs' results strengthen the hypothesis that cultural distance has significant negative influences on international trade.

#### 6.4.2. Additional result at the country level

As a robust test, we apply our GEE estimates to country-level data using two different grouping strategies detailed in Section 6.3. The estimation results presented by Table 7 show that for both types of grouping, the proposed GEE approach continues to demonstrate the negative impacts of both geographical distance and cultural distance on trade, as well as the positive effect of the GDP per capita in the gravity equation. For the main explanatory variable, the cultural distance, the Poisson GEE estimated coefficients for trade, export, and import in the first type of grouping are −0.407, −0.346, and −0.519. Overall, the estimation results from country-level data under two alternative grouping strategies are comparable to those from product-level data, showing only moderate differences in the coefficient magnitudes.

**Table 7.** Poisson GEE for country-level data.

	Trade Grouping 1	Export Grouping 1	Import Grouping 1	Trade Grouping 2	Export Grouping 2	Import Grouping 2
logCD	−0.407***	−0.346***	−0.519**	−0.401**	−0.305*	−0.505*
s.e.	(0.109)	(0.077)	(0.162)	(0.143)	(0.120)	(0.241)
logGDPPC	0.900**	0.840*	1.008***	0.776***	0.698**	0.869***
s.e.	(0.346)	(0.424)	(0.238)	(0.230)	(0.264)	(0.237)
logDist	−0.522*	−0.521	−0.535**	−0.644*	−0.610*	−0.719*
s.e.	(0.219)	(0.266)	(0.202)	(0.290)	(0.306)	(0.283)
Language	−0.897*	−0.293	−2.045***	−0.587	0.033	−1.646
s.e.	(0.407)	(0.383)	(0.450)	(0.461)	(0.354)	(1.015)
Landlock	−1.198***	−1.682***	−1.003***	−1.778***	−2.358***	−1.424**
s.e.	(0.291)	(0.279)	(0.258)	(0.425)	(0.534)	(0.454)
constant	14.625**	14.402*	13.073***	16.663***	16.356**	15.763***
s.e.	(5.000)	(6.230)	(3.615)	(4.390)	(4.986)	(3.838)
N	95	95	95	95	95	95

s.e. stands for standard errors that are robust to model misspecification.

## 7. Conclusion and further work

We propose a GEE method to estimate nonlinear models in the presence of spatially dependent innovations. To target applications with latent causal links, we focus on nonlinear models with spatial errors. We suggest grouping the data to adjust for the dependence induced by the spatial errors, with a grouped working variance-covariance matrix accounting for the within-group dependence. We list a condition on the working variance-covariance matrix under which the proposed spatial GEE estimator has the actual efficiency gain relative to the ungrouped QMLE estimator, and show that the former does perform better than the latter in simulations for various data-generating processes. The specific estimation procedures of the proposed method are given for the Probit binary model and the Poisson count model, and the asymptotic properties and a consistent estimator of the variance-covariance matrix of the proposed GEE estimator are provided. In the end, to illustrate the usage of our method, we implement the proposed GEE method to the extended gravity equation with the trade data between China and the rest of the world, and document the important role of the cultural distance on international trade.

We suggest the following directions for further research. First, throughout the article, we assume that the group structure is exogenous. We do not consider the case where groups are endogenously generated. One can investigate further how to model endogenous group accounting for a class of general nonlinear models with spatial dependence. Second, although in many applications, such as Section 6, natural grouped information is available to specify the working variance-covariance matrix, there are cases where this information is not available. How to specify a group structure in the first step within the current framework deserves further attention. Third, as pointed out by one of the referees, extending the current method to models where the dependent variables are also spatially correlated would also be an important and interesting future research topic.

## Acknowledgments

We would like to express our most sincere thanks to the editor, the associate editor, and two anonymous reviewers for very valuable comments and suggestions.

## Data availability statement

The data that support the findings of this study are available from the author, Cuicui Lu (lucucui@sdu.edu.cn), upon reasonable request.

## Disclosure statement

The authors report there are no competing interests to declare.

## Funding

Cuicui Lu gratefully acknowledges the support from National Social Science Foundation of China (Grant No. 20BJY017). Weining Wang's research is partially supported by the ESRC (Grant No. ES/T01573X/1).

## References

- Adegboye, O. A., Leung, D. H. Y., Wang, Y.-G. (2018). Analysis of spatial data with a nested correlation structure. *Journal of the Royal Statistical Society Series C: Applied Statistics* 67(2):329–354. doi: [10.1111/rssc.12230](https://doi.org/10.1111/rssc.12230)
- Anderson, J. E., Van Wincoop, E. (2003). Gravity with gravitas: A solution to the border puzzle. *American Economic Review* 93(1):170–192. doi: [10.1257/000282803321455214](https://doi.org/10.1257/000282803321455214)
- Angrist, J. D. (2014). The perils of peer effects. *Labour Economics* 30:98–108. doi: [10.1016/j.labeco.2014.05.008](https://doi.org/10.1016/j.labeco.2014.05.008)
- Arbia, G. (2016). Spatial econometrics: A rapidly evolving discipline. *Econometrics*, 4(1):1–4.
- Arbia, G., Billé, A. G. (2018). *Spatial discrete choice models: A review focused on specification, estimation and health economics applications*. BEMPS - Bozen Economics & Management Paper Series, BEMPS54, Faculty of Economics and Management at the Free University of Bozen.
- Baltagi, B. H., Egger, P. H., Kesina, M. (2016). Bayesian spatial bivariate panel probit estimation. In: *Spatial Econometrics: Qualitative and Limited Dependent Variables*, Vol. 37. Emerald Publishing Ltd., pp. 119–144.
- Baltagi, B. H., Song, S. H., Koh, W. (2003). Testing panel data regression models with spatial error correlation. *Journal of Econometrics* 117(1):123–150. doi: [10.1016/S0304-4076\(03\)00120-9](https://doi.org/10.1016/S0304-4076(03)00120-9)
- Bloom, N., Schankerman, M., Van Reenen, J. (2013). Identifying technology spillovers and product market rivalry. *Econometrica*, 81(4):1347–1393.
- Brown, J. R., Ivković, Z., Smith, P. A., Weisbenner, S. (2008). Neighbors matter: Causal community effects and stock market participation. *The Journal of Finance* 63(3):1509–1531. doi: [10.1111/j.1540-6261.2008.01364.x](https://doi.org/10.1111/j.1540-6261.2008.01364.x)
- Cameron, A. C., Trivedi, P. K. (1986). Econometric models based on count data. Comparisons AND applications of some estimators and tests. *Journal of Applied Econometrics* 1(1):29–53. doi: [10.1002/jae.3950010104](https://doi.org/10.1002/jae.3950010104)
- Carrell, S. E., Sacerdote, B. I., West, J. E. (2013). From natural variation to optimal policy? The importance of endogenous peer group formation. *Econometrica* 81(3):855–882.
- Carter, A. V., Schnepel, K. T., Steigerwald, D. G. (2017). Asymptotic behavior of a t-test robust to cluster heterogeneity. *The Review of Economics and Statistics* 99(4):698–709. doi: [10.1162/REST\\_a\\_00639](https://doi.org/10.1162/REST_a_00639)
- Conley, T. G. (1999). GMM estimation with cross sectional dependence. *Journal of Econometrics* 92(1):1–45. doi: [10.1016/S0304-4076\(98\)00084-0](https://doi.org/10.1016/S0304-4076(98)00084-0)
- Cressie, N. (2015). *Statistics for Spatial Data*. New York: John Wiley & Sons.
- Davidson, J. (2021). *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford: Oxford University Press.
- De Paula, Á. (2013). Econometric analysis of games with multiple equilibria. *Annual Review of Economics* 5(1):107–131. doi: [10.1146/annurev-economics-081612-185944](https://doi.org/10.1146/annurev-economics-081612-185944)
- Elhorst, J. P. (2014). Spatial panel data models. In: *Spatial Econometrics*. Heidelberg: Springer, pp. 37–93.
- Gagliardini, P., E., Ossola, O., Scaillet, (2020). Estimation of large dimensional conditional factor models in finance, *Handbook of Econometrics*, 7, 219, 282: Elsevier.
- Gourieroux, C., Monfort, A., Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica* 52(3):681. doi: [10.2307/1913471](https://doi.org/10.2307/1913471)
- Graham, B. S. (2008). Identifying social interactions through conditional variance restrictions. *Econometrica* 76(3):643–660. doi: [10.1111/j.1468-0262.2008.00850.x](https://doi.org/10.1111/j.1468-0262.2008.00850.x)
- Hofstede, G., Hofstede, G. J., Minkov, M. (2010). *Cultures and Organizations: Software of the Mind*. New York: McGraw-Hill.
- Jacod, J., Sørensen, M. (2018). A review of asymptotic theory of estimating functions. *Statistical Inference for Stochastic Processes* 21(2):415–434. doi: [10.1007/s11203-018-9178-8](https://doi.org/10.1007/s11203-018-9178-8)
- Jenish, N., Prucha, I. R. (2009). Central limit theorems and uniform laws of large numbers for arrays of random fields. *Journal of Econometrics* 150(1):86–98. doi: [10.1016/j.jeconom.2009.02.009](https://doi.org/10.1016/j.jeconom.2009.02.009)
- Jenish, N., Prucha, I. R. (2012). On spatial processes and asymptotic inference under near-epoch dependence. *Journal of Econometrics* 170(1):178–190. doi: [10.1016/j.jeconom.2012.05.022](https://doi.org/10.1016/j.jeconom.2012.05.022)
- Kapoor, M., Kelejian, H. H., Prucha, I. R. (2007). Panel data models with spatially correlated error components. *Journal of Econometrics* 140(1):97–130. doi: [10.1016/j.jeconom.2006.09.004](https://doi.org/10.1016/j.jeconom.2006.09.004)
- Kelejian, H. H., Prucha, I. R. (2007). HAC estimation in a spatial framework. *Journal of Econometrics* 140(1):131–154. doi: [10.1016/j.jeconom.2006.09.005](https://doi.org/10.1016/j.jeconom.2006.09.005)

- Krueger, A. B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics* 114(2):497–532. doi: [10.1162/003355399556052](https://doi.org/10.1162/003355399556052)
- Lee, L.-F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica* 72(6):1899–1925. doi: [10.1111/j.1468-0262.2004.00558.x](https://doi.org/10.1111/j.1468-0262.2004.00558.x)
- Lee, L.-F., Liu, X., Lin, XU. (2010). Specification and estimation of social interaction models with network structures. *Econometrics Journal* 13(2):145–176. doi: [10.1111/j.1368-423X.2010.00310.x](https://doi.org/10.1111/j.1368-423X.2010.00310.x)
- Lesage, J., Pace, R. K. (2009). *Introduction to Spatial Econometrics*. New York: Chapman and Hall/CRC.
- Liang, K.-Y., Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73(1):13–22. doi: [10.1093/biomet/73.1.13](https://doi.org/10.1093/biomet/73.1.13)
- Lin, P.-S., Clayton, M. K. (2005). Analysis of binary spatial data by quasi-likelihood estimating equations. *The Annals of Statistics*, 33(2):542–555.
- Newey, W. K., Mcfadden, D. (1994). Large sample estimation and hypothesis testing. In: *Handbook of Econometrics*, Vol. 4. Amsterdam: Elsevier, pp. 2111–2245.
- Newey, W. K., Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica* 71(5):1565–1578. doi: [10.1111/1468-0262.00459](https://doi.org/10.1111/1468-0262.00459)
- Oman, S. D., Landsman, V., Carmel, Y., Kadmon, R. (2007). Analyzing spatially distributed binary data using independent-block estimating equations. *Biometrics* 63(3):892–900. doi: [10.1111/j.1541-0420.2007.00754.x](https://doi.org/10.1111/j.1541-0420.2007.00754.x)
- Pinkse, J., Slade, M. E. (1998). Contracting in space: An application of spatial statistics to discrete-choice models. *Journal of Econometrics* 85(1):125–154. doi: [10.1016/S0304-4076\(97\)00097-3](https://doi.org/10.1016/S0304-4076(97)00097-3)
- Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* 44(4):1033. doi: [10.2307/2531733](https://doi.org/10.2307/2531733)
- Rao Chaganty, N., Joe, H. (2004). Efficiency of generalized estimating equations for binary responses. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 66(4):851–860. doi: [10.1111/j.1467-9868.2004.05741.x](https://doi.org/10.1111/j.1467-9868.2004.05741.x)
- Sacerdote, B. (2001). Peer effects with random assignment: Results for dartmouth roommates. *The Quarterly Journal of Economics* 116(2):681–704. doi: [10.1162/00335530151144131](https://doi.org/10.1162/00335530151144131)
- Silva, J. M. C. S., Tenreiro, S. (2006). The log of gravity. *The Review of Economics and Statistics* 88(4):641–658. doi: [10.1162/rest.88.4.641](https://doi.org/10.1162/rest.88.4.641)
- Tinbergen, J. (1962). An analysis of world trade flows. *Shaping the World Economy* 3:1–117.
- Van Der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge: Cambridge University Press.
- Wang, H., Iglesias, E. M., Wooldridge, J. M. (2013). Partial maximum likelihood estimation of spatial probit models. *Journal of Econometrics* 172(1):77–89. doi: [10.1016/j.jeconom.2012.08.005](https://doi.org/10.1016/j.jeconom.2012.08.005)
- Wooldridge, J. M. (2003). Cluster-sample methods in applied econometrics. *American Economic Review* 93(2):133–138. doi: [10.1257/000282803321946930](https://doi.org/10.1257/000282803321946930)
- Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.
- Wooldridge, J. M. (2020). *Introductory Econometrics: A Modern Approach*. Mason, OH: Cengage Learning.
- Xu, X., Lee, L.-F. (2015a). Maximum likelihood estimation of a spatial autoregressive tobit model. *Journal of Econometrics* 188(1):264–280. doi: [10.1016/j.jeconom.2015.05.004](https://doi.org/10.1016/j.jeconom.2015.05.004)
- Xu, X., Lee, L.-F. (2015b). A spatial autoregressive model with a nonlinear transformation of the dependent variable. *Journal of Econometrics* 186(1):1–18. doi: [10.1016/j.jeconom.2014.12.005](https://doi.org/10.1016/j.jeconom.2014.12.005)