DSSR

Discussion Paper No. 132

Spatial Extension of the Mixed Models of the Analysis of Variance

Takaki Sato , Yuta Kuroda,
and Yasumasa Matsuda

October, 2022

Data Science and Service Research
Discussion Paper

Center for Data Science and Service Research
Graduate School of Economic and Management
Tohoku University
27-1 Kawauchi, Aobaku
Sendai 980-8576, JAPAN

Spatial Extension of the Mixed Models of the Analysis of Variance

Takaki Sato *[1], Yuta Kuroda[2], and Yasumasa Matsuda[3]

[1]Economic Department, Musashi University, Tokyo, Japan

[2]Graduate School of Economics, Osaka Metropolitan University, Osaka, Japan

[3]Graduate School of Economics and Management, Tohoku University, Sendai, Japan

Abstract

This paper proposes a spatial extension of the mixed models of the analysis of variance(MANOVA) models, which are called mixed spatial ANOVA (MS-ANOVA) models. MS-ANOVA models have been used to evaluate spatial correlations between random effects in multilevel data which is a kind of cluster data in which observations belong to some kinds of nested clusters. Because the proposed model can be regarded as a Bayesian hierarchical models, we have introduced empirical Bayesian estimation methods in which hyper parameters are estimated by quasi-maximum likelihood estimation methods in the first step and posterior distributions for the parameters are evaluated with the estimated hyper-parameters in the second step. Moreover, we have justified the asymptotic properties of the first step estimators. The proposed models are applied to happiness survey data in Japan and empirical results show that social capital which can be interpreted as "the beliefs and norms by which a community values collective action and pursues activities worthy for the entire community" significantly increases people's happiness, even after controlling for a variety of individual characteristics and spatial correlations.

Keywords: Spatial econometrics, Multilevel data, ANOVA model, Empirical Bayesian estimation, Quasi-maximum likelihood estimation.

1 Introduction

This paper aims to extend mixed analysis of variance (MANOVA) models for multilevel data (see Demidenko (2013)) to those for spatial multilevel data, which we call spatial error models for multilevel data (MS-

*Corresponding author. E-mail address: takaki.sato@cc.musashi.ac.jp

ANOVA). Multilevel data is a kind of cluster data in which observations belong to some kinds of nested clusters (e.g. students are members of one of schools in school effectiveness research) and widely used in both of social and natural science (see De Leeuw et al. (2008) and Hox et al. (2017)). Spatial multilevel data in which clusters are organized with spatial regions are also used in many fields to capture spatial correlation between regions. With spatial multilevel data, Fazio & Piacentino (2010) investigates spatial variability of small and medium enterprises productivity across the Italian territory, and Pierewan & Tampubolon (2014) examines how spatial clusters explain variations in individual well-begin across regions in Europe.

Mixed models of the analysis of variance (MANOVA) models which are linear regression models incorporating several kinds of random effect terms corresponding to cluster types have been used to analyze multilevel data. By including random effect terms, we can evaluate common feature within same clusters as grouping structures. Asymptotic properties of maximum likelihood estimator for MANOVA models are discussed in Hartley and Rao (1967) and Miller (1977).

To evaluate spatial correlations between random effects in multilevel data, we provide a spatial extension of MANOVA models, which are called mixed spatial ANOVA (MS-ANOVA) models in this paper. The conventional way to estimate spatial correlation is to include spatial lag terms into regression models (see Anselin (1988) and Arbia (2014)), and thus we combine spatial lag terms with random effect terms in MANOVA models to propose new spatial econometrics models. Because MS-ANOVA models are defined by two level equations which are individual level and regional level equations, the models can be regarded as hierarchical bayesian models whose parameters and hyperparameters can be estimated by empirical bayesian estimation methods in two steps. In the first step, hyperparameters are estimated with quasi-maximum likelihood (QML) estimation methods which are common methods in spatial econometrics studies (see Lee (2004), Liu and Yang (2015), Su and Yang (2015), and Yang (2018)) Posterior distributions for parameters are derived with hyperparemters estimated in the first step.

The interesting feature of MS-ANOVA models are summarized as follows. Firstly, MS-ANOVA models can analyze regional specific effects for dependent variables, considering the effect of individual characteristics. Here, let us note that regional effects are not the same as random effects. The regional effect of one region are defined by the effect of observed regional characteristics and the sum of random effects for clusters, each of which corresponds to the group in clusters which the region belongs to. Spatial econometrics models that have been considered ever can't take into account of the effect of individual characteristics in estimating regional effect in multilevel data because we need to summarize the data on regions where more than one individuals are observed to apply conventional models, and then individual characteristics are lost. Defining MS-ANOVA models in two level equations would allow that we can take into account both individual characteristics and regional effects in analysis. Secondly, spatial correlations between random effects can be estimated. Some sources of random effects such as cultures or customs specific to a region may tend to be similar to them in nearby regions. Then, random effects may have spatial correlation, namely, regional effects in nearby areas may take on similar values. Therefore, taking into account the spatial correlation between random effects improves the accuracy of the model fitting. Thirdly, we can estimate regional effects in areas where there are no observed individuals by using the information of the region where observed individuals exist. Because regional effects are estimated by Bayesian estimation methods, we can evaluate regional effects for all regions, regardless of whether the individuals belong to them or not by properly estimating the prior information of regional effects.

Applications of MS-ANOVA models to happiness survey data in Japan demonstrate several interesting features of the effect of individual characteristics and regional specific characteristics on happiness. Firstly, individual characteristics are important factor for happiness. People's happiness is U-shaped with respect to age, namely, happiness decreases until middle age and then increases. Moreover, female is basically happier than male. Happiness increases monotonically as household income and personal income increase and getting married greatly increases people's happiness. Secondly, random effects for each city have spatial correlation. The similarity of culture or customs of nearby areas which greatly affects the way people feel about their happiness may cause the spatial correlation. Finally, spatial cluster exists in the regional effects of Japanese happiness survey data which can be regarded as average happiness of each regions. The level of happiness in the southern and middle regions of Japan is higher than that in the eastern region, and the estimated happiness of eastern coastal regions is the lowest. The reason is thought to be that the effects of the nuclear accident caused by the East-Japan earthquake which occurred in 2011 are still lingering.

This paper is organized as follow. In Section 2, we define MS-ANOVA models as a spatial extension of MANOVA models. A two step estimation method to estimate the parameters in MS-ANOVA models and asymptotic properties of the first step estimator are discussed in Section 3. We apply the MS-ANOVA model to happiness survey data in Japan to demonstrate empirical properties of the proposed models in Section 4. Section 5 concludes the paper. All the proofs in Section 3 are discussed in Appendix.

2 Model specification

Let n and L be the number of individual and regions, respectively. We assume that each individuals belong to one of the regions and admit that there are regions where no individuals are observed. In this paper, we call the nested dataset whose grouping is based on spatial units as spatial multilevel data. Moreover, we assume that the regions can be grouped into larger regional units by p different groupings, and let $m_l, l = 1, \ldots, p$, be the number of larger regions obtained by the l-th grouping. For examples, several cities are grouped together to form a prefecture in Japan.

Suppose that spatial multilevel data $y_{i,j}$ is the dependent variable for the i-th individual belonging to the j-th region and Y is the $n \times 1$vector of $y_{i,j}$s. MS-ANOVA models are given by,

$$Y = X_1\beta_1 + Jd + \varepsilon, \tag{1}$$

$$d = X_2\beta_2 + u, \tag{2}$$

$$u = U_1(I_1 - \rho_1 W_1)^{-1} f_1 + \cdots U_p(I_p - \rho_p W_p)^{-1} f_p, \tag{3}$$

where X_1 is an $n \times k_1$ matrix for individual level explanatory variables, J is an $n \times L$ matrix for regional dummy variables, X_2 is an $L \times k_2$ matrix for regional level explanatory variables, U_l is an $L \times m_l$ matrix for a random effect which consists only of zeros and ones and there is exactly one 1 in each row and at least one 1 in each column, $l = 1, \ldots, p$, I_l is an $m_l \times m_l$ identity matrix, and W_l is an $m_l \times m_l$ spatial weight matrix which describes spatial relationships among the l-th grouped regions. A random variables $\varepsilon_i, i = 1, \ldots, n$, is independent and identically distributed (i.i.d.) with mean 0 and variance σ_0^2 and an $n \times 1$ vector $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$, and $f_{l,j}, j = 1, \ldots, m_l$ is also i.i.d. with mean 0 and variance σ_l^2 and an $m_l \times 1$ vector $f_l = (f_{l1}, \ldots, f_{lm_l})'$ is a random effect for the l-the groped regions. The vector β_1 and β_2 are regression coefficients for individual and regional level explanatory variables, respectively, and ρ_l is a spatial correlation parameter which describe the strength of spatial dependence between regions in the l-th grouping.

MS-ANOVA models are a spatial extension of MANOVA models because MS-ANOVA models reduce to MANOVA models when spatial correlation parameters, ρ_ls are equal to 0. In the analysis of spatial multilevel data, consideration of spatial correlation between random effects can improve the accuracy of the model fitting. Random effects $f_{l,j}$ express the effect of the j-th region in the l-th larger regional units on regional effect, d, and some sources of random effects may be cultures or customs in the larger region. Because the cultures and customs of nearby regions tend to be similar, random effects $f_{l,j}$ may have spatial correlation, namely, regional effects in nearby areas may take on similar values. Therefore, taking into account the spatial correlation between regional effects allows for more detailed analysis of the spatial multilevel data.

MS-ANOVA models are defined by two level equations, individual level equations (1) and regional level equations (2) and (3), and this two level modeling has some advantages. One advantage is that we can analyze regional effect, d, considering the effect of individual characteristics. Spatial econometrics models that have been considered can't take into account of individual characteristics when we estimate regional effect in multilevel data because they assume that exactly one observation is observed each regions. Thus, we need to summarize the multilevel data which may have more than one observations in each regions so that there is one value for one region. A commonly used method to summarize data is to take the average of the data in the area, but then individual characteristics are lost then. Defining the MS-ANOVA models in two level equations would allow for both individual and regional effects, which would allow for more accurate estimation of regional effects.

Another advantage of modeling multilevel models with the hierarchical structure is that we can estimate regional effects, d, in areas where there are no observed individuals by using the information of the regions where observed individuals exist. Let us remember that J is a regional dummy matrix which may have columns whose elements are all zeros because we admit the existence of regions where no individuals are observed. Thus, usual ordinary least squares does not work because J is rank deficient. The proposed model can be regarded as a Bayesian hierarchical model and equation (2) and (3) describe prior information of regional effects. As discussed in the estimation section, by properly estimating the prior information of regional effects with marginal likelihood which is based on the information of regions where individuals are observed, we can evaluate regional effects for all regions, regardless of whether the individuals belong to them or not.

3 Estimation

Let us consider a method to estimate the parameters for MS-ANOVA models and discuss asymptotic properties of proposed estimators the size of $m_l, l = 1, \ldots, p$, tends to be infinity along with the sample size n. Because the proposed model can be regarded as a Bayesian hierarchical model, we propose empirical Bayes estimation procedure in two steps. The first step is the estimation of the hyperparameters in prior distributions with quasi-maximum likelihood (QML) estimation methods, and the second step is calculation of posterior distributions with the hyperparameters estimated in the first step. Moreover, we introduce the asymptotic properties of the first step estimators when the sample size of both individuals and regions tends to be large.

3.1 Empirical Bayes Estimation

We introduce empirical Bayes estimation procedure in two steps. Let $\beta = (\beta_1, \beta_2)$, $\tau_l = \frac{\sigma_l^2}{\sigma_0^2}, l = 1, \ldots, p$, $\theta = (\beta_2', \tau_1, \ldots, \tau_p, \rho_1, \ldots, \rho_p)'$, $\psi = (\beta', \sigma_0^2, \ldots, \sigma_p^2, \rho_1, \ldots, \rho_p)'$ and $\delta = (\beta_1, d)$. In this paper, we call σ_0^2 and θ as hyperparameters and δ as parameters, respectively.

The first step is the estimation of σ_0^2 and θ by a quasi-maximum likelihood (QML) estimation method with a marginal likelihood of Y. Let us denote that $f(Y|\beta_1, d, \sigma_0^2)$ is a probability density function for the data Y and $g(d|\beta_2, \rho_1, \ldots, \rho_p, \sigma_1^2, \ldots, \sigma_p^2)$ is a prior distribution of the variable d. In this step, we regard random variables ε_i and $f_{l,j}$ which may be not normally distributed random variables follows normal distribution, and then the marginal distribution of Y,

$$m(Y|\psi) = \int f(Y|\beta, \sigma_\varepsilon^2, d) g(d|\beta_2, \rho_1, \ldots, \rho_p, \sigma_1^2, \ldots, \sigma_p^2) \boldsymbol{d} d,$$

follows a multivariate normal distribution. Thus, the marginal log-likelihood function is given by

$$\log L(\psi) = -\frac{n}{2} \log(2\pi\sigma_0^2) - \frac{1}{2} \log |\Omega(\theta)| - \frac{(Y - X\beta)'\Omega^{-1}(\theta)(Y - X\beta)}{2\sigma_0^2}, \tag{4}$$

where $X = (X_1, JX_2)$, $\Omega(\theta) = I_n + \tau_1 JU_1(I_1 - \rho_1 W_1)^{-1}(I_1 - \rho_1 W')^{-1}U_1 J' + \cdots + \tau_p JU_p(I_p - \rho_p W_p)^{-1}(I_p - \rho_p W_p')^{-1}U_p J'$.

We derive a concentrated marginal log-likelihood function to reduce the number of parameters for numerical optimization. The first-order condition of the marginal log-likelihood function is

$$\hat{\beta}(\theta) = (X'\Omega^{-1}(\theta)X)^{-1}X'\Omega^{-1}(\theta)Y,$$
$$\hat{\sigma}_0^2(\theta) = \frac{1}{n}(Y - X\hat{\beta}(\theta))'\Omega^{-1}(\theta)(Y - X\hat{\beta}(\theta)).$$

By substituting $\hat{\beta}(\theta)$ and $\hat{\sigma}_0^2(\theta)$ into (4), we obtain the concentrated marginal log-likelihood function,

$$\log L(\theta) = -\frac{n}{2}(\log(2\pi) + 1) - \frac{n}{2} \log(\hat{\sigma}_0^2(\theta)) - \frac{1}{2} \log |\Omega(\theta)|.$$

Maximizing the concentrated marginal log-likelihood function gives the QML estimator $\hat{\theta}$ of θ, and then the QML estimators $\hat{\beta}$ and $\hat{\sigma}_0^2$ are obtained by $\hat{\beta} = \hat{\beta}^*(\hat{\theta})$ and $\hat{\sigma}_0^2 = \hat{\sigma}_0^2(\hat{\theta})$, respectively.

The second step is the Bayesian estimation of the parameters δ based on the estimated hyperparameters $\hat{\beta}_2$, $\hat{\rho}$, $\hat{\sigma}_0^2$ and $\hat{\sigma}_l^2 = \hat{\tau}_l \hat{\sigma}_0^2, l = 1, \ldots, p$. Let $\delta = (\beta', d')'$ and $\tilde{X} = (X, J)$. The estimated posterior distribution for δ is given by

$$P(\delta \mid Y, \tilde{X}, \hat{\beta}_2, \hat{\rho}_1, \ldots, \hat{\rho}_p, \hat{\sigma}_0^2, \ldots, \hat{p}_f^2, b) \propto L(Y|\tilde{X}, \delta, \hat{\sigma}_0^2)\pi(\delta|\hat{\beta}_2, \hat{\rho}, \hat{\sigma}_1^2, \ldots, \hat{\sigma}_p^2, b),$$

where b is a hyperparameter for a prior information for β, $L(Y|\tilde{X}, \theta, \hat{\sigma}_0^2)$ is the likelihood of the data Y, and $\pi(\theta|\hat{\beta}_2, \hat{\rho}, \hat{\sigma}_1^2, \ldots, \hat{\sigma}_p^2, b)$ is the prior distribution of the model parameters, δ. If the prior and posterior distribution for δ is conjugate distributions, then prior distribution can be calculated explicit form, and if not conjugate distributions, samples of δ from posterior distribution are obtained by Markov chain Monte Carlo (MCMC) methods.

As one example of conjugate distributions which is used in the empirical application session in this papaer, we will show the explicit form of the posterior distribution when the likelihood and the prior distribution are multivariate normal distributions and the number of random effect is one. Then, the estimated posterior distribution follows a multivariate normal distribution. We set prior means and the inverse of prior variance matrices of the multivariate normal distribution for the prior distribution as $\hat{s}_0 = (0'_{k\times 1}, \hat{\beta}_2{}')'$ and $\hat{S}_0^{-1} = \begin{pmatrix} 0_{k\times k} & 0_{k\times m} \\ 0_{m\times k} & \frac{1}{\hat{\sigma}_1^2}(I_n - \hat{\rho}W')(I_n - \hat{\rho}W) \end{pmatrix}$, where $0_{n_1\times n_2}$ is the $n_1 \times n_2$ matrix whose elements are zeros. Then, the posterior covariance matrix and mean vector is $S_1 = \left(\frac{1}{\hat{\sigma}_0^2}\tilde{X}'\tilde{X} + \hat{S}_0^{-1}\right)^{-1}$ and $s_1 = S_1\left(\frac{1}{\hat{\sigma}_0^2}\tilde{X}'Y + \hat{S}_0^{-1}\hat{s}_0\right)$, respectively. Thus, the estimated posterior distribution is given by

$$P(\delta \mid Y, \tilde{X}, \hat{\beta}_2, \hat{\rho}_1, \hat{\sigma}_0^2, \hat{\sigma}_1^2, b) \sim N(s_1, S_1),$$

where $N(s_1, S_1)$ means the multivariate normal distribution with mean s_1 and covariance matrix S_1.

3.2 Asymptotic properties

We discuss the conditions under which the QML estimators $\hat{\theta} = (\hat{\beta}_2', \hat{\tau}_1, \ldots, \hat{\tau}_p, \hat{\rho}_1, \ldots, \hat{\rho}_p)'$ and $\hat{\psi} = (\hat{\beta}', \hat{\sigma}_0^2, \ldots, \hat{\sigma}_p^2, \hat{\rho}_1, \ldots, \hat{\rho}_p)'$ in the first step is consistent and asymptotically normal when the size of $m_l, l = 1, \ldots, p$, tends to be infinity along with the sample size n. All of the proofs and Lemmas for the asymptotic results are given in the Appendix.

Let $\theta_0 = (\delta_0', \tau_{10}, \ldots, \tau_{p0}, \rho_{10}, \ldots, \rho_{p0})'$ and $\psi_0 = (\beta_0', \sigma_{00}^2, \theta_0')'$be the true values for θ and ψ. Assume the following conditions.

Assumption 1 The true parameter θ_0 lies in the interior of a compact parameter space Θ.

Asuumption 2 $\varepsilon_i, i = 1, \ldots, n$ and $f_{l,j}, l = 1, \ldots, p, j = 1, \ldots, m_l$ are i.i.d with mean 0 and variances σ_0^2 and σ_j^2, respectively. And, $E|\varepsilon_i|^{4+\delta} < \infty$ and $E|f_{l,j}|^{4+\delta} < \infty$ for some $\delta > 0$.

Assumption 3 The number of regions in l-th grouping, m_l, tends to infinity along with the sample size n.

Assumption 4 The matrices J, U_i, W_i, $(I_j - \rho_{j,0}W_j)^{-1}$ and $\Omega^{-1}(\theta)$ is uniformly bounded in both row and column sums. Moreover, $0 < \underline{c}_\omega \leq \inf_{\theta\in\Theta} \gamma_{\min}(\Omega^{-1}(\theta)) \leq \sup_{\theta\in\Theta} \gamma_{\max}(\Omega^{-1}(\theta)) \leq \bar{c}_\omega < \infty$.

Assumption 5 X has full column rank and its elements are uniformly bounded constants, $\lim_{n\to\infty} \frac{1}{n} X'\Omega^{-1}(\theta)X$ exists and is non-singular.

Assumption 6 Let $A_i^{-1}(\rho_i) = (I_i - \rho_{i,0}W_j)^{-1}(I_i - \rho_{i,0}W_i)^{-1}$ and $B_i(\rho_i) = (I_i - \rho_i W_i)^{-1}W_i' + W_i(I_i - \rho_i W_i')^{-1}$. We assume that $\sup_{\theta\in\Theta} |\gamma_{\max}(JU_iA_i^{-1}(\rho_i)U_i'J')| < \infty$ and $\sup_{\theta\in\Theta} |\gamma_{\max}(JU_iA_i^{-1}(\rho_i)B_i(\rho_i)A_i^{-1}(\rho_i)U_i'J')| < \infty, i = 1, \ldots, p$.

Assumption 7

$$\lim_{n\to\infty} \frac{1}{n}\left[\log|\sigma_{00}^2\Omega(\theta_0)| - \log|\tilde{\sigma}^2(\theta)\Omega(\theta)|\right] \neq 0, \text{for any } \theta \neq \theta_0.$$

First, we introduce the consistency of $\hat{\theta}$. The expected log-likelihood function for the proposed model is given by

$$E\log L(\psi) = -\frac{n}{2}\log(2\pi\sigma_0^2) - \frac{1}{2}\log|\Omega(\theta)| - E\left(\frac{(Y - X\beta)'\Omega^{-1}(\theta)(Y - X\beta)}{2\sigma_0^2}\right).$$

The expected log-likelihood is maximized at

$$\begin{aligned}
\tilde{\beta}(\theta) &= \beta_0, \\
\tilde{\sigma}_0^2(\theta) &= \frac{\sigma_{00}^2}{n} tr(\Omega(\theta)^{-1}\Omega(\theta_0)), \\
&= \frac{1}{n}E[u_0'\Omega^{-\frac{1}{2}}(\theta)M(\theta)\Omega^{-\frac{1}{2}}(\theta)u_0'] + \frac{1}{n}E[u_0'\Omega^{-\frac{1}{2}}(\theta)P(\theta)\Omega^{-\frac{1}{2}}(\theta)u_0'],
\end{aligned}$$

where $P(\theta) = I - M(\theta)$ and $M(\theta) = I - \Omega^{-\frac{1}{2}}(\theta)X(X'\Omega^{-1}(\theta)X)^{-1}X'\Omega^{-\frac{1}{2}}(\theta)$. Thus, the concentrated expected log-likelihood function is given by,

$$E \log L(\theta) = -\frac{n}{2}(\log(2\pi) + 1) - \frac{n}{2}\log(\tilde{\sigma}_0^2(\theta)) - \frac{1}{2}\log|\Omega(\theta)|.$$

Consistency of $\hat{\theta}$ is obtained by the following two facts. The first one is the identification uniqueness condition: $\limsup_{n\to\infty}\left[\max_{\theta\in B^c(\theta_0,\varepsilon)\cap\Theta} E\log L(\theta) - E\log L(\theta_0)\right] < 0$ for any $\varepsilon > 0$, where $B^c(\theta_0,\varepsilon)$ is the compliment of an ε-neighborhood of θ_0. The second one is the uniform convergence in probability: $\sup_{\theta\in\Theta}\left|\frac{1}{n}\log L(\theta) - \frac{1}{n}E\log L(\theta)\right| = o_p(1)$.

Theorem 1. *Under Assumptions 1-7, $\hat{\theta}$ is a consistent estimator of θ_0.*

Next, let us consider the the asymptotic distribution of the QML estimator $\hat{\psi}$. To derive the asymptotic normality, we need to consider the the Taylor expansion of $\frac{\partial}{\partial\psi}\log L_n(\hat{\psi})$ at ψ_0. The first-order derivatives of the log-likelihood function at ψ has the elements

$$\begin{aligned}
\frac{\partial \log L(\psi)}{\partial \beta} &= X'\Sigma^{-1}(\eta)(Y - X\beta),\\
\frac{\partial \log L(\psi)}{\partial \sigma_i^2} &= -\frac{1}{2}tr(\Sigma^{-1}(\eta)G_i(\rho_i)) + \frac{1}{2}(Y - X\beta)'\Sigma^{-1}(\eta)G_i(\rho_i)\Sigma^{-1}(\eta)(Y - X\beta),\\
\frac{\partial \log L(\theta)}{\partial \rho_i} &= \frac{\sigma_i^2}{2}tr(\Sigma^{-1}(\eta)H_i(\rho_i)) - \frac{\sigma_i^2}{2}(Y - X\beta)'\Sigma^{-1}(\eta)H_i(\rho_i)\Sigma^{-1}(\eta)(Y - X\beta),
\end{aligned}$$

where $\eta = (\sigma_0^2, \sigma_1^2, \ldots, \sigma_p^2, \rho_1, \ldots, \rho_p)$, $A_i^{-1}(\rho_i) = (I_i - \rho_i W_i)^{-1}(I_i - \rho_i W_i')^{-1}$, $B_i(\rho_i) = W_i'(I_i - \rho_i W_i) + (I_i - \rho_i W_i')W_i$, $G_i(\rho_i) = JU_iA_i^{-1}(\rho_i)U_i'J'$, and $H_i(\rho_i) = JU_iA_i^{-1}(\rho_i)B_i(\rho_i)A_i^{-1}(\rho_i)U_i'J'$. By the mean value theorem,

$$\sqrt{n}(\hat{\psi} - \psi_0) = -\left(\frac{1}{n}\frac{\partial^2 \log L(\bar{\psi})}{\partial\psi\partial\psi'}\right)^{-1}\frac{1}{\sqrt{n}}\frac{\partial \log L(\psi_0)}{\partial\psi},$$

where $\bar{\psi}$ lies between $\hat{\psi}$ and ψ_0.

The score function which is the first-order derivatives of the log-likelihood function at ψ_0, $\frac{\partial \log L(\psi_0)}{\partial\psi}$, are linear and quadratic functions of $u_0 = (Y - X\beta_0)$. By applying the central limit theorem for linear-quadratic forms by Kelejian and Prucha (2001) to the score functions, we have the asymptotic normality for the QMLE $\hat{\psi}$ under proper asymptotic behavior of the Hessian matrix and the variance of the score function whose explicit forms are given in Appendix.

Theorem 2. *Under Assumptions 1-7, if there exist* $\Sigma = -\lim_{n\to\infty} E\left(\frac{1}{n}\frac{\partial^2 \log L(\psi_0)}{\partial\psi\partial\psi'}\right)$

and $\Gamma = \lim_{n\to\infty} \frac{1}{n} E\left(\frac{\partial \log L(\psi)}{\partial \psi}\frac{\partial \log L(\psi)}{\partial \psi'}\right)$, *and* $-\Sigma$ *is positive definite, then*

$$\sqrt{n}(\hat{\psi} - \psi_0) \xrightarrow{D} N(0, \Sigma^{-1}\Gamma\Sigma^{-1}).$$

4 Empirical Application

We conduct empirical analysis for MS-ANOVA models by real data analysis for happiness survey data in Japan to analyze the effect of individual characteristics on happiness and spatial correlation between regional effects. Moreover, we investigate the relationship between happiness and social capital which can be interpreted as "the beliefs and norms by which a community values collective action and pursues activities worthy for the entire community".

In recent years, happiness has been extensively investigated in the social sciences. Various studies have revealed relationships between several individual characteristics and happiness. Oswald (1997) reported a U-shaped relationship between age and happiness, that is, the happienss of people in middle age is the smallest among all age groups. In addition, women tend to have higher average happiness at all ages than men. Helliwell (2003) reports that, after controlling for many individual factors, people who are single are less happy than those who are married. However, there has been little discussion on geographical feature of happiness. Regional factors such as culture and customs are believed to have a significant impact on people's happiness. If spatial correlations exist for such unobservable regional factors, an analysis that ignores spatial correlations is likely to lead to erroneous conclusions. By using the proposed model, this analysis will take into account the spatial correlation of such unobservable regional factors and clarify the geographical characteristics of happiness.

Social capital can be interpreted as "the beliefs and norms by which a community values collective action and pursues activities worthy for the entire community" and is an indicator related to spatial correlates of hapiness (Bartscher et al. 2021; Durante et al., 2021). Activities that are privately costly and have no direct reward, but are socially useful, such as voting (Bauernschuster et al., 2014; Guiso et al., 2004), and blood and organ donation (Buonanno et al., 2009), are often used as a proxy for social capital. Communities with high social capital provide more public goods and services because people are more mutually beneficial and trusting, and are more cooperative in achieving the common purposes of the community. Therefore, hapiness could be increased in areas with high social capital through deeper friendships with neighbors, increased local communal activities and informal helps, abundant and diverse provision of public amenities, improved security and public health, and reduced unlawful activities such as fraud and corruption.

Let us introduce the happiness survey data. In December 2019, we commissioned Macromill Co, LTD, which is a market research company in Japan, to conduct a survey of 26, 974 people living in 1534 cities. Here, respondents were selected so that the distribution of age, population, and area of residence would be the same as that of the Japanese census. The demographic information of the respondents contains gender, age, personal and family incomes, marital status, jobs and presence of children.

Happiness for dependent variables, Y, was obtained by asking individuals to answer the following question: *Currently, how happy do you feel? Score the degree of your happiness between 1 (very unhappy) and 10 (very happy).* Thus, the happiness is measured discrete values between 1 and 10.

We use dummy variables created from the demographic information as explanatory variables, X. Age and gender are divided into 12 categories, namely, all the respondents were separated into the two groups of female and male, each of which is categorized as the 6 mutually disjoint subgroups corresponding with 10s, 20s, 30s, 40s, 50s, and over 60. The group of female in their 20s as the base group. Personal income is categorized as the 9 mutually disjoint groups of income, *i.e.* (1) $<$ 2 million yen, (2) $<$ 4 million, (3) $<$ 6 million, (4) $<$ 8 million, (5) $<$ 10 million, (6) $<$ 12 million (7) $<$ 15 million, (8) $<$ 20 million, and (9) $\geq$ 20 million yen, with the group less than 2 million yen as the base. Moreover, the grouping of household income is based on the grouping of personal income plus the group of no-response. The base group for household income is the group less than 2 million yen Presence of child is summarized as the dummy variable of taking 1 if a respondent has more than one child and 0 otherwise. Martial status is recorded as the category variable with the three groups of (1) single, (2) married and (3) divorced or widowed, with the single group taken as the base.

In addition, regional level explanatory variables, X_2, are voting rates, percentage of population over 65 years old. We use voting rates which is a social capital proxy variables to clear the relationship between people's happiness and local social capital. Other variables are used to control for local economic conditions.

Regional dummy variables, J, is the $26,974 \times 1838$ matrix. Here, we note that the number of all municipalities in Japan is 1845 and the rank of J is less than 1838, at 1534 which is the number of cities which at least one respondent belong to. Let us remember that our proposed models can also estimate regional effects, d, in areas where there are no observed individuals by using the information of the surrounding regions where observed individuals exist. Thus, the matrix, J, contains the columns whose elements are zeros which correspond to the areas where there are no respondents.

We use a 1838×1838 spatial weight matrix, W, created in two-steps. Firstly, If the distance between

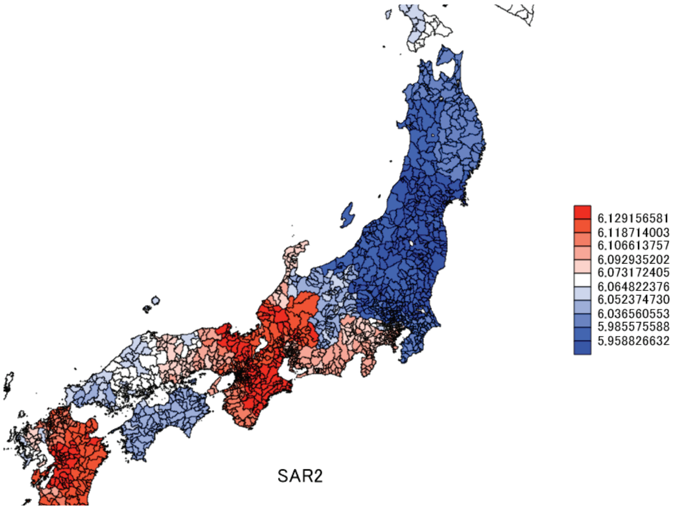

Figure 1: a plot of estimated regional effects for individual happiness for each cities with MS-ANOVA models which contains city level and prefecture level random effects by applying it them to happiness dataset in Japan.

city i and j is 30 km, the i, j and j, i elements of W is 1, and otherwise 0. Next, if there are no more than three cities within 30 km of a city, then the element of W corresponding to the three closest cities is set to 1. After that each row sum is standardized to be 1.

Table 1 reports the estimates of the parameters for MS-ANOVA models with standard errors and Akaike information criterion (AIC) for both models. As a benchmark for comparison, those of MANOVA models which is a special case of MS-ANOVA models where spatial parameters for random effects are equal to zeros, i.e. $\rho = 0$. In comparison between fittings of MS-ANOVA models and MANOVA models, the former model accounts for happiness better than the later model in terms of AIC. This indicated that taking into account of spatial correlation improve the accuracy of the model fitting. The results regarding the relationship between individual characteristics and happiness are similar to those of existing studies. For example, people in middle age have the lowest level of happiness, and in each age group, women's happiness is higher than men's happiness. Moreover, happiness increases monotonically as household income and personal income increase and getting married greatly increases people's happiness.

Next, let us consider the spatial correlates of happiness. We find from table 1 that spatial correlation between random effects are positively significant at 5 % level, which indicates that random effect on a city takes similar value with random effects on surrounding cities of the city. One reason which derives spatial correlation between city level random coefficient is the similarity of culture or customs which greatly affects the way people feel about their happiness. Figure 1 is a plot of estimated regional effects for each cities, d, which mean average happiness of people living in that city. We can find that spatial cluster in the regional effects exists and the level of happiness in the southern and middle regions of Japan is higher than that in the eastern region. Especially, eastern coastal regions show the lowest level of happiness compared to other regions. The reason for this is thought to be that the effects of the nuclear accident caused by the East-Japan earthquake which occurred in 2011 are still lingering.

We examine the relationship between the level of social capital in an area and the level of happiness of the people living in the area. Table1 shows that the estimate of voting rates, a proxy measure of social capital in the area, is statistically significant at 0.631. We find results consistent with previous studies (e.g. Hommerich & Tiefenbach, 2018) that social capital of residence significantly increases people's happiness, even after controlling for a variety of individual characteristics and spatial correlations. Since social capital is a composite indicator, a more detailed analysis is needed to determine which aspects of it drive the results, and this is a fruitful subject for the future.

5 Conclusion

We have proposed MS-ANOVA models which is a spatial extension of the mixed models of the analysis of variance in this paper. Because the proposed model can be regarded as a Bayesian hierarchal model, we have introduced empirical bayesian estimation methods in two steps as estimation strategy for the parameters in the proposed models. The first step estimator specifies the hyper parameters and has been justified in asymptotic situations, and the second step estimator for parameters are derived by the Bayes' formula with the hyperparameters estimated in the first step. Fitting the proposed model to happiness survey data in Japan, we can evaluate the effect of individual and regional level explanatory variables on happiness and spatial correlation of regional effects which are random effects in each regions. Empirical results suggest that happiness is U-shaped with age, female's happiness is higher than male's happiness at all ages, and regional effects on happiness are spatially correlated. The existence of spatial correlations between random effects indicates that unobserved features which affect on individual happiness such as culture and custom tend to be similar in nearby regions.

For future study, several extensions are possible. In this analysis, we regard individual happiness as continuous variables. However, the treatment creates a gap between the data and the model because individual happiness takes only discrete values between 1 and 10. Thus, the extension of the proposed model to discrete choice models fills the gap and allows for rigorous analysis of happiness. One more possibility is a panel extension of the proposed model. Our proposed model can capture only spatial correlation. However, it is said that happiness on individual has a time series correlation. A panel extension would reveal more interesting sptio-temporal correlation in happiness.

References

[1] Andrews, D. W. (1992). Generic uniform convergence. *Econometric theory, 8*(2), 241-257.

[2] Anselin, L. (1988). *Spatial econometrics: methods and models.* Netherlands: Kluwer.

[3] Arbia, G. (2014). *A primer for spatial econometrics: with applications in R.* Springer.

[4] Bartscher, A. K., Sebastian S., Sebastian S., Michaela S., & Nils W. (2021). Social capital and the spread of covid-19: Insights from european countries, *Journal of Health Economics 80*, 102531.

[5] Bauernschuster, S., Oliver F., & Ludger W. (2014). Surfing alone? The internet and social capital: Evidence from an unforeseeable technological mistake. *Journal of Public Economics 117*, 73–89.

[6] Buonanno, P., Daniel M., & Paolo V. (2009). Does social capital reduce crime? *The Journal of Law and Economics 52*(1), 145–170.

[7] Davidson, J. (1994). *Stochastic limit theory.* Oxford: Oxford University Press.

[8] Demidenko, E. (2013). *Mixed models: theory and applications with R.* John Wiley & Sons.

[9] Fazio, G., & Piacentino, D. (2010). A spatial multilevel analysis of Italian SMEs' productivity. *Spatial Economic Analysis, 5*(3), 299-316.

[10] Durante, R., Luigi G., & Giorgio G. (2021). Asocial capital: Civic culture and social distancing during COVID-19. *Journal of Public Economics 194* (2021), 104342.

[11] Guiso, L., Paola S., & Luigi Z. (2004). The role of social capital in financial development. *American Economic Review 94*(3), 526–556.

[12] Hommerich, C., & Tiefenbach, T. (2018). Analyzing the Relationship Between Social Capital and Subjective Well-Being: The Mediating Role of Social Affiliation. *Journal of Happiness Studies, 19*(4), 1091-1114.

[13] Hartley, H. O., & Rao, J. N. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika, 54*(1), 93-108.

[14] Hellisell, J. F. (2003). How's life? combining individual and national variables to explain subjective well-being. *Economic modeling 20* (2), 331-360.

[15] Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications.* Routledge.

[16] De Leeuw, J., Meijer, E., & Goldstein, H. (2008). *Handbook of multilevel analysis.* New York: Springer.

[17] Kelejian, H. H., & Prucha, I. R. (2001). On the asymptotic distribution of the Moran I test statistic with applications. *Journal of econometrics, 104*(2), 219-257.

[18] Lee, L. F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica, 72*(6), 1899-1925.

[19] Liu, S. F., & Yang, Z. (2015). Asymptotic distribution and finite sample bias correction of QML estimators for spatial error dependence model. *Econometrics, 3*(2), 376-411.

[20] Miller, J. J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *The Annals of Statistics, 5*(4), 746-762.

[21] Oswald, A. J. (1997). Happiness and economic performance. *The economic journal, 107*(445), 1815-1831.

[22] Pierewan, A. C., & Tampubolon, G. (2014). Spatial dependence multilevel model of well-being across regions in Europe. *Applied Geography, 47*, 168-176.

[23] Su, L., & Yang, Z. (2015). QML estimation of dynamic panel data models with spatial errors. *Journal of econometrics, 185*(1), 230-258.

[24] White, H. (1994). *Estimation, inference and specification analysis.* New York: Cambridge university press.

[25] Yang, Z. (2018). Unified M-estimation of fixed-effects spatial dynamic models with short panels. *Journal of econometrics, 205*(2), 423-447.

Table 1: Estimation values and their standard errors for β_1, β_2, log likelihood (logL) and Akaike Information Criterion (AIC) in both MS-ANOVa models and mixed analysis of variance models (MANOVA) which contains city level and prefecture level random effects, and estimates and standard errors of spatial parameters ρ_1 and ρ_2 in MS-ANOVA models which are obtained by applying them to happiness dataset in Japan.

	MS-ANOVA		MANOVA	
	coef	s.e	coef	s.e
20< Female < 25	-0.007	0.076	-0.005	0.076
Female < 30	-0.234	0.071	-0.232	0.071
Female < 35	-0.522	0.069	-0.521	0.069
Female < 40	-0.694	0.068	-0.691	0.068
Female < 45	-0.718	0.064	-0.717	0.064
Female < 50	-0.857	0.064	-0.856	0.064
Female < 55	-0.811	0.067	-0.813	0.066
Female < 60	-0.789	0.068	-0.790	0.068
Female < 65	-0.504	0.067	-0.506	0.067
Female > 65	-0.172	0.063	-0.172	0.062
Male < 20	0.185	0.078	0.184	0.078
Male < 25	-0.186	0.072	-0.183	0.072
Male < 30	-0.836	0.070	-0.835	0.070
Male < 35	-1.117	0.068	-1.118	0.068
Male < 40	-1.409	0.069	-1.411	0.068
Male < 45	-1.557	0.065	-1.558	0.065
Male < 50	-1.628	0.065	-1.629	0.065
Male < 55	-1.692	0.068	-1.694	0.068
Male < 60	-1.799	0.069	-1.802	0.069
Male < 65	-1.308	0.068	-1.311	0.068
Male < 65	-0.915	0.065	-0.917	0.064
200 < Personal Income (PI) < 400	0.076	0.032	0.075	0.032
PI < 600	0.263	0.041	0.264	0.041
PI < 800	0.335	0.057	0.338	0.057
PI < 1000	0.401	0.081	0.403	0.081
PI < 1200	0.572	0.120	0.577	0.120
PI > 1200	0.458	0.146	0.459	0.146
200 < Family Income (FI) < 400	0.055	0.048	0.055	0.048
FI < 600	0.337	0.048	0.339	0.048
FI < 800	0.494	0.052	0.495	0.052
FI < 1000	0.619	0.058	0.621	0.058
FI < 1200	0.769	0.069	0.770	0.069
FI < 1500	0.881	0.087	0.882	0.087
FI < 2000	1.181	0.117	1.183	0.117
FI > 2000	1.111	0.162	1.109	0.163
FI Unknown	0.137	0.045	0.138	0.045
Married	0.961	0.041	0.961	0.041
Divorced	0.390	0.057	0.391	0.057
Children	0.109	0.035	0.110	0.035
voting rates	0.631	0.266	0.600	1.303
population over 65 years	-0.002	0.003	-0.006	0.013
rho1(city)	0.440	0.013		
rho2(Pref)	0.252	1.217		
logL	-1.2050		-1.2055	
AIC	60164		60193	

Appendix A. Hessian and Covariance matrix

Here, we show the detailed expression of the Hessian matrix and covariance matrix which is discussed in Theorem 2. Firstly, we show the Hessian matrix. For simplicity, we denote $A_i^{-1}(\rho_i) = (I_i - \rho_i W_i)^{-1}(I_i - \rho_i W_i')^{-1}$, $B_i(\rho_i) = W_i'(I_i - \rho_i W_i) + (I_i - \rho_i W_i')W_i$, $G_i(\rho_i) = JU_iA_i^{-1}(\rho_i)U_i'J'$, $H_i(\rho_i) = JU_iA_i^{-1}(\rho_i)B_i(\rho_i)A_i^{-1}(\rho_i)U_i'J'$, $H_{1,i}(\rho_i) = JU_iA_i^{-1}(\rho_i)B_i(\rho_i)A_i^{-1}(\rho_i)B_i(\rho_i)A_i^{-1}(\rho_i)U_i'J'$, and $H_{2,i}(\rho_i) = JU_iA_i^{-1}(\rho_i)W_i'W_iA_i^{-1}(\rho_i)U_i'J', i = 1, \ldots, p$. Moreover, we define $A_0(\rho_0)^{-1} = I_L$ and $U_i = I_L$. Then, the variance matrix of the proposed model is given by, $\Sigma(\eta) = \sum_{i=0}^{p} \sigma_i^2 G_i(\rho_i)$, where $\eta = (\sigma_0^2, \sigma_1^2, \ldots, \sigma_p^2, \rho_1, \ldots, \rho_p)$. Moreover, the derivatives of $G_i(\rho_i), H_i(\rho_i), \Sigma(\eta)$ are given by, $\frac{\partial G_i(\rho_i)}{\partial \rho_i^2} = -H_i(\rho_i)$, $\frac{\partial H_i(\rho_i)}{\partial \sigma_i^2} = -2(H_{1,i}(\rho_i) + H_{2,i}(\rho_i))$, $\frac{\partial \Sigma(\eta)}{\partial \sigma_i^2} = G_i(\rho_i)$ and $\frac{\partial \Sigma(\eta)}{\partial \rho_i} = -\sigma_i^2 H_i(\rho_i)$, respectively.

By using above notations, the gradients of the log-likelihood function, $\frac{\partial \log L(\psi)}{\partial \psi}$, is given by

$$\begin{aligned}
\frac{\partial \log L(\psi)}{\partial \beta} &= X'\Sigma^{-1}(\eta)(Y - X\beta), \\
\frac{\partial \log L(\psi)}{\partial \sigma_i^2} &= -\frac{1}{2} tr(\Sigma^{-1}(\eta)G_i(\rho_i)) + \frac{1}{2}(Y - X\beta)'\Sigma^{-1}(\eta)G_i(\rho_i)\Sigma^{-1}(\eta)(Y - X\beta), \\
\frac{\partial \log L(\theta)}{\partial \rho_i} &= \frac{\sigma_i^2}{2} tr(\Sigma^{-1}(\eta)H_i(\rho_i)) - \frac{\sigma_i^2}{2}(Y - X\beta)'\Sigma^{-1}(\eta)H_i(\rho_i)\Sigma^{-1}(\eta)(Y - X\beta).
\end{aligned}$$

Moreover, the hessian matrix of the log-likelihood function, $\frac{\partial^2 \log L(\psi)}{\partial \psi \partial \psi'}$ has the elements:

$$\begin{aligned}
\frac{\partial^2 \log L(\psi)}{\partial \beta \partial \beta'} &= -X\Sigma^{-1}(\eta)X, \\
\frac{\partial^2 \log L(\psi)}{\partial \beta \partial \sigma_i^2} &= -X'\Sigma^{-1}(\eta)G_i(\rho_i)\Sigma^{-1}(\eta)(Y - X\beta), \\
\frac{\partial^2 \log L(\psi)}{\partial \beta \partial \rho_i} &= \sigma_i^2 X'\Sigma^{-1}(\eta)H_i(\rho_i)\Sigma^{-1}(\eta)(Y - X\beta), \\
\frac{\partial^2 \log L(\psi)}{\partial \sigma_i^2 \partial \sigma_i^2} &= \frac{1}{2} tr(\Sigma^{-1}(\eta)G_i(\rho_i)\Sigma^{-1}(\eta)G_i(\rho_i)) \\
&\quad - (Y - X\beta)'\Sigma^{-1}(\eta)G_i(\rho_i)\Sigma^{-1}(\eta)G_i(\rho_i)\Sigma^{-1}(\eta)(Y - X\beta),
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial^2 \log L(\psi)}{\partial \sigma_i^2 \partial \sigma_j^2} &= \frac{1}{2} tr(\Sigma^{-1}(\eta) G_j(\rho_j) \Sigma^{-1}(\eta) G_i(\rho_i)) \\
&\quad - (Y - X\beta)' \Sigma^{-1}(\eta) G_j(\rho_j) \Sigma^{-1}(\eta) G_i(\rho_i) \Sigma^{-1}(\eta)(Y - X\beta), \\
\frac{\partial^2 \log L(\psi)}{\partial \sigma_i^2 \partial \rho_i} &= \frac{\sigma_i^2}{2} tr(\Sigma^{-1}(\eta) H_i(\rho_i) \Sigma^{-1}(\eta) G_i(\rho_i)) - \frac{1}{2} tr(\Sigma^{-1}(\eta) H_i(\rho_i)) \\
&\quad + \sigma_i^2 (Y - X\beta)' \Sigma^{-1}(\eta) H_i(\rho_i) \Sigma^{-1}(\eta) G_i(\rho_i) \Sigma^{-1}(\eta)(Y - X\beta) \\
&\quad - \frac{1}{2} (Y - X\beta)' \Sigma^{-1}(\eta) H_i(\rho_i) \Sigma^{-1}(\eta)(Y - X\beta) \\
\frac{\partial^2 \log L(\psi)}{\partial \sigma_i^2 \partial \rho_j} &= -\frac{\sigma_j^2}{2} tr(\Sigma^{-1}(\eta) H_j(\rho_j) \Sigma^{-1}(\eta) G_i(\rho_i)) \\
&\quad + \sigma_j^2 (Y - X\beta)' \Sigma^{-1}(\eta) H_j(\rho_j) \Sigma^{-1}(\eta) G_i(\rho_i) \Sigma^{-1}(\eta)(Y - X\beta), \\
\frac{\partial^2 \log L(\psi)}{\partial \rho_i \partial \rho_i} &= \frac{\sigma_i^4}{2} tr(\Sigma^{-1}(\eta) H_i(\rho_i) \Sigma^{-1}(\eta) H_i(\rho_i)) - \sigma_i^2 tr(\Sigma^{-1}(\eta)(H_{1,i}(\rho_i) + H_{2,i}(\rho_i)) \\
&\quad - \frac{\sigma_i^4}{2} (Y - X\beta)' \Sigma^{-1}(\eta) H_i(\rho_i) \Sigma^{-1}(\eta) H_i(\rho_i) \Sigma^{-1}(\eta)(Y - X\beta) \\
&\quad + \sigma_i^2 (Y - X\beta)' \Sigma^{-1}(\eta)(H_{1,i}(\rho_i) + H_{2,i}(\rho_i)) \Sigma^{-1}(\eta)(Y - X\beta), \\
\frac{\partial^2 \log L(\psi)}{\partial \rho_i \partial \rho_j} &= \frac{\sigma_i^2 \sigma_j^2}{2} tr(\Sigma^{-1}(\eta) H_j(\rho_j) \Sigma^{-1}(\eta) H_i(\rho_i)) \\
&\quad - \frac{\sigma_i^2 \sigma_j^2}{2} (Y - X\beta)' \Sigma^{-1}(\eta) H_j(\rho_j) \Sigma^{-1}(\eta) H_i(\rho_i) \Sigma^{-1}(\eta)(Y - X\beta).
\end{aligned}
$$

Next, let us consider the variance matrix of the log likelihood function, $E(\frac{\partial \log L(\psi_0)}{\partial \psi} \frac{\partial \log L(\psi_0)}{\partial \psi'})$. The explicit form of each elements can be obtained form Lemma 4 in Appendix B:

$$
\begin{aligned}
E\left(\frac{\partial \log L(\psi_0)}{\partial \beta} \frac{\partial \log L(\psi_0)}{\partial \beta'}\right) &= X' \Sigma^{-1}(\eta_0) X, \\
E\left(\frac{\partial \log L(\psi_0)}{\partial \beta} \frac{\partial \log L(\psi_0)}{\partial \sigma_i^2}\right) &= \frac{1}{2} X' \Sigma^{-1}(\eta_0) E(u_0 u_0' \Sigma^{-1}(\eta_0) G_i(\rho_{0i}) \Sigma^{-1}(\eta_0) u_0), \\
E\left(\frac{\partial \log L(\psi_0)}{\partial \beta} \frac{\partial \log L(\psi_0)}{\partial \rho_i}\right) &= -\frac{\sigma_{0i}^2}{2} X' \Sigma^{-1}(\eta_0) E(u_0 u_0' \Sigma^{-1}(\eta_0) H_i(\rho_{0i}) \Sigma^{-1}(\eta_0) u_0),
\end{aligned}
$$

$$
\begin{aligned}
E\left(\frac{\partial \log L(\psi_0)}{\partial \sigma_i^2}\frac{\partial \log L(\psi_0)}{\partial \sigma_i^2}\right) &= -\frac{1}{4}[E(u_0'\Sigma^{-1}(\eta_0)G_i(\rho_{0i})\Sigma^{-1}(\eta_0)u_0)]^2 + \frac{1}{4}E[(u_0'\Sigma^{-1}(\eta_0)G_i(\rho_{0i})\Sigma^{-1}(\eta_0)u_0)^2],\\
E\left(\frac{\partial \log L(\psi_0)}{\partial \sigma_i^2}\frac{\partial \log L(\psi_0)}{\partial \sigma_j^2}\right) &= -\frac{1}{4}E(u_0'\Sigma^{-1}(\eta_0)G_i(\rho_{0i})\Sigma^{-1}(\eta_0)u_0)E(u_0'\Sigma^{-1}(\eta_0)G_j(\rho_{0j})\Sigma^{-1}(\eta_0)u_0)\\
&\quad + \frac{1}{4}E[u_0'\Sigma^{-1}(\eta_0)G_i(\rho_{0i})\Sigma^{-1}(\eta_0)u_0u_0'\Sigma^{-1}(\eta_0)G_j(\rho_{0j})\Sigma^{-1}(\eta_0)u_0],\\
E\left(\frac{\partial \log L(\psi_0)}{\partial \sigma_i^2}\frac{\partial \log L(\psi_0)}{\partial \rho_j}\right) &= \frac{\sigma_j^2}{4}E(u_0'\Sigma^{-1}(\eta_0)G_i(\rho_{0i})\Sigma^{-1}(\eta_0)u_0)E(u_0'\Sigma^{-1}(\eta_0)H_j(\rho_{0j})\Sigma^{-1}(\eta_0)u_0),\\
&\quad - \frac{1}{4}E[u_0'\Sigma^{-1}(\eta_0)G_i(\rho_{0i})\Sigma^{-1}(\eta_0)u_0u_0'\Sigma^{-1}(\eta_0)H_j(\rho_{0j})\Sigma^{-1}(\eta_0)u_0],\\
E\left(\frac{\partial \log L(\psi_0)}{\partial \rho_i^2}\frac{\partial \log L(\psi_0)}{\partial \rho_j}\right) &= -\frac{\sigma_j^2}{4}E(u_0'\Sigma^{-1}(\eta_0)H_i(\rho_{0i})\Sigma^{-1}(\eta_0)u_0)E(u_0'\Sigma^{-1}(\eta_0)H_j(\rho_{0j})\Sigma^{-1}(\eta_0)u_0),\\
&\quad + \frac{1}{4}E[u_0'\Sigma^{-1}(\eta_0)H_i(\rho_{0i})\Sigma^{-1}(\eta_0)u_0u_0'\Sigma^{-1}(\eta_0)H_j(\rho_{0j})\Sigma^{-1}(\eta_0)u_0],
\end{aligned}
$$

where $\eta_0 = (\sigma_{00}^2, \sigma_{01}^2, \dots, \sigma_{0p}^2, \rho_{01}, \dots, \rho_{0p})$.

6 Appendix B. Some useful lemmas

We introduce some lemmas which are used in the proofs of the following main results. The lemmas are a little modifications of lemmas in Lee(2004) for non-square matrices.

Lemma 1 Let A be an $n \times m$ non-square matrix whose column sums are uniformly bounded, C be a $n \times k$ matrix whose elements are uniformly bounded, and f_i be i.i.d noise with mean 0 and variance σ^2. Then, $\frac{1}{\sqrt{m}}C'Af = O_p(1)$.

Proof. Let $B = C'A$, $b_{i,j}$ be the (i, j)-th element of B and b_i be the i-th counn of B. Because the elements of C are uniformly bounded and the column sums of A are uniformly bounded, the element of B is uniformly bounded by Lemmas in Lee (2004). Let $\bar{b}$ be a constant such as $|b_{i,j}| \leq \bar{b}$. Because $Bf = \sum_{i=1}^m b_i f_i$, $Var(Bf) = E(\sum_{i=1}^m \sum_{i=1}^m b_i f_i f_j b_j') = \sigma^2 \sum_{i=1}^m b_i b_i' \leq \sum_{i=1}^m \bar{b}1_k 1_k' = O(m)$, where 1_k is a $k \times 1$ vector whose elements are 1. Thus, $\frac{1}{\sqrt{m}}C'Af = O_p(1)$ by Chebyshev's inequality. □

Lemma 2 Let A be an $m_1 \times m_2$ non-square matrix whose column sums are uniformly bounded, $f_{1,i}$ and $f_{2,i}$ are i.i.d noise with mean 0 and variance σ_1 and σ_2, respectively. Then,

- $E(f_1 A f_2) = 0$.
- $V(f_1 A f_2) = O(m_1)$.

- $f_1 A f_2 = O_p(\sqrt{m_1})$.

Proof. $E(f_1Af_2) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} a_{i,j}E(f_{1,i}f_{2,j}) = 0$. $V(f_1Af_2) =$ $\sum_{i_1=1}^{m_1}\sum_{i_2=1}^{m_1}\sum_{j_1=1}^{m_2}\sum_{j_2=1}^{m_2} a_{i_1,j_1}a_{i_2,j_2}E(f_{1,i_1}f_{1,j_2}f_{2,j_1}f_{2,j_2}) = \sigma_1^2\sigma_2^2\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} a_{i,j}^2 \leq \sigma_1^2\sigma_2^2\sum_{i=1}^{m_1}(\sum_{j=1}^{n}|a_{i,j}|)^2 \leq$ $\sigma_1^2\sigma_2^2\sum_{i=1}^{m_1} c^2 = O(m_1)$. Thus, $f_1Af_2 = O_p(\sqrt{m_1})$ by Chebyshev's inequality. $\square$

Lemma 3 Let A_i be an $m_i \times m_i$ matrix for $i = 1, \ldots, p$, B be an $n \times n$ matrix, C be an $n \times k$ matrix, and ε and $f_i, i = 1, \ldots, p$ be an $n \times 1$ and $m_i \times 1$ random noise with means 0 and variances σ_0^2 and σ_i^2. Moreover, we define U_i be an $n \times m_i$ matrix which consists only of zeros and ones and there exist one 1 in each row and at least one 1 in each column, $i = 1, \ldots, p$ We denote $u = \varepsilon + \sum_{i=1}^{p} U_iA_if_i$ and $m = \min\{m_1, \ldots, m_p\}$. We assume U_i is uniformly bounded in column sums, the elements of C is uniformly bounded, B is uniformly bounded in both row and column sums, and m_i is a function of n and tends to infinity and $\lim_{n\to\infty}\frac{m_i}{n} = c_i \leq 1$. Then,

- $\frac{1}{n}C'Bu = o_p(1)$.
- $\frac{1}{n}u'Bu = O_p(1)$.
- $\frac{1}{n}(u'Bu - E(u'Bu)) = o_p(1)$.

Proof. By the Lemma,

$$\begin{aligned}
\frac{1}{n}C'Bu &= \frac{1}{n}C'B\varepsilon + \sum_{i=1}^{p}\frac{1}{n}C'BU_iA_if_i,\\
&= \frac{1}{\sqrt{n}}\frac{1}{\sqrt{n}}C'B\varepsilon + \sum_{i=1}^{p}\frac{m_i}{n}\frac{1}{\sqrt{m_i}}\frac{1}{\sqrt{m_i}}C'BU_iA_if_i,\\
&= o(1)O_p(1) + \sum_{i=1}^{p}O(1)o(1)O_p(1),\\
&= o_p(1).
\end{aligned}$$

We denote $f_0 = \varepsilon_0, A_0 = I_n$ and $U_0 = I_n$. Because $u_0 = U_0A_0f_0 + U_1A_1f_1 + \cdots + U_pA_pf_p$,

$$\frac{1}{m}u'Bu = \sum_{i=0}^{p}\sum_{j=0}^{p}\frac{1}{m}f_i'A_i'U_i'BU_jA_jf_j.$$

Firstly, we will consider the case When i = j. Because $A_i'U_i'BU_iA_i$ is uniformly bounded in both row and column sums, by Lee(2004),

$$\frac{1}{n}f_i'A_i'U_i'BU_iA_if_i = \frac{m_i}{n}\frac{1}{m_i}f_i'A_i'U_i'BU_iA_if_i,$$
$$= O(1)O_p(1),$$
$$= O_p(1).$$

Moreover, $\frac{m_i}{n}\frac{1}{m_i}(f_i'A_i'U_i'BU_iA_if_i - E(f_i'A_i'U_i'BU_iA_if_i) = O(1)o_p(1) = o_p(1)$.

Secondly, we will consider the case of $i \neq j$. By the Lemma,

$$\frac{1}{n}f_i'A_i'U_i'BU_jA_jf_j = \frac{m_i}{n}\frac{1}{\sqrt{m_i}}\frac{1}{\sqrt{m_i}}f_i'A_i'U_i'BU_jA_jf_j,$$
$$= O(1)o(1)O_p(1),$$
$$= o_p(1).$$

It is clear that $\frac{m_i}{n}\frac{1}{m_i}(f_i'A_i'\frac{m_i}{n}U_i'\frac{n}{m}BU_jA_jf_j - E(f_i'A_i'\frac{m_i}{n}U_i'\frac{n}{m}BU_jA_jf_j) = o_p(1)$

Therefore, $\frac{1}{n}u'Bu = O_p(1)$ and $\frac{1}{n}(u'Bu - E(u'Bu)) = o_p(1)$. □

Lemma 4 Let A be an $m_1 \times m_2$ non-square matrix, $T_i = JU_i(I_i - \rho_iW_i)^{-1}$, and f_i $i = 0, \ldots, p$ are $m_i \times 1$ i.i.d. random noise with mean 0 and variances σ_i^2, respectively. Moreover, the elements of each f_i has more than fourth moment, i.e. $E|f_{1,i}|^{4+\delta} < \infty$ for some $\delta > 0$. Let us define $u = \sum_{i=1}^p T_if_i$. Then,

1. $E(uu') = \Sigma(\eta)$.
2. $E(u'Au) = \sum_{i=0}^p \sigma_i^2 tr(T_i'AT_i)$.
3. $E(uu'Au) = \sum_{i=0}^p \mu_{i,3}T_i diag(T_iAT_i)$.
4. $E(u'Auu'Bu) = \sum_{i=1}^p \Big[(\mu_{i,4} - 3\sigma_i^4)\sum_{j=1}^n (T_i'AT_i)_{j,j}(T_i'BT_i)_{j,j} + \sigma_i^4(tr(T_i'AT_i)tr(T_i'BT_i) + tr(T_i'AT_i(T_i'(B+B')T_i))\Big] + \sum_{i_1}^p\sum_{i_2}^p \sigma_{i_1}^2\sigma_{i_2}^2 tr(T_{i_1}'AT_{i_1})tr(T_{i_2}'BT_{i_2}) + 2\sum_{i_1}^p\sum_{i_2}^p \sigma_{i_1}^2\sigma_{i_2}^2 tr(T_{i_1}'AT_{i_1}T_{i_2}'B'T_{i_2})$

Proof.

$$
\begin{aligned}
E(uu') &= E\Big(\sum_{i=0}^{p} T_i f_i\Big)\Big(\sum_{i=0}^{p} f_i' T_i'\Big),\\
&= \sum_{i=0}^{p} T_i E(f_i f_i') T_i',\\
&= \sum_{i=0}^{p} \sigma_i^2 T_i T_i',\\
&= \Sigma(\eta).\\
E(u'Au) &= E\Big(\sum_{i_1=0}^{p}\sum_{i_2=0}^{p} f_{i_1}' T_{i_1}' A T_{i_2} f_{i_2}\Big),\\
&= \sum_{i=0}^{p} E(f_i' T_i' A T_i f_i),\\
&= \sum_{i=0}^{p} \sigma_i^2 tr(T_i' A T_i),\\
E(uu'Au) &= E\Big(\sum_{i_1=0}^{p}\sum_{i_2=0}^{p}\sum_{i_3=0}^{p} T_{i_1} f_{i_1} f_{i_2}' T_{i_2}' A T_{i_3} f_{i_3}\Big),\\
&= \sum_{i=0}^{p} E(T_i f_i f_i' T_i A T_i f_i),\\
&= \sum_{i=0}^{p} E\Big(T_i f_i \sum_{j_1=1}^{m_i}\sum_{j_2=1}^{m_i} (T_i A T_i)_{j_1,j_2} f_{i,j_1} f_{i,j_2}\Big),\\
&= \sum_{i=0}^{p} \mu_{i,3} T_i diag(T_i A T_i).\\
E(u'Auu'Bu) &= E\Big(\sum_{i_1}^{p}\sum_{i_2}^{p}\sum_{i_3}^{p}\sum_{i_4}^{p} f_{i_1}' T_{i_1}' A T_{i_2} f_{i_2} f_{i_3}' T_{i_3}' B T_{i_4} f_{i_4}\Big),\\
&= \sum_{i}^{p} E(f_i' T_i' A T_i' f_i f_i' T_i' B T_i' f_i) + \sum_{i_1}^{p}\sum_{i_2\neq i_1}^{p} E(f_{i_1}' T_{i_1}' A T_{i_1} f_{i_1} f_{i_2}' T_{i_2}' B T_{i_2} f_{i_2})\\
&+ \sum_{i_1}^{p}\sum_{i_2\neq i_1}^{p} E(f_{i_1}' T_{i_1}' A T_{i_2} f_{i_2} f_{i_1}' T_{i_1}' B T_{i_2} f_{i_2}) + \sum_{i_1}^{p}\sum_{i_2\neq i_1}^{p} E(f_{i_1}' T_{i_1}' A T_{i_2} f_{i_2} f_{i_2}' T_{i_2}' B T_{i_1} f_{i_1}),\\
&= \sum_{i=1}^{p}\Big[(\mu_{i,4} - 3\sigma_i^4)\sum_{j=1}^{n} (T_i' A T_i)_{j,j} (T_i' B T_i)_{j,j} + \sigma_i^4\big(tr(T_i' A T_i) tr(T_i' B T_i) + tr(T_i' A T_i (T_i'(B+B')T_i))\big)\Big]\\
&+ \sum_{i_1}^{p}\sum_{i_2}^{p} \sigma_{i_1}^2 \sigma_{i_2}^2 tr(T_{i_1}' A T_{i_1}) tr(T_{i_2}' B T_{i_2}) + 2\sum_{i_1}^{p}\sum_{i_2}^{p} \sigma_{i_1}^2 \sigma_{i_2}^2 tr(T_{i_1}' A T_{i_1} T_{i_2}' B' T_{i_2})
\end{aligned}
$$

□

Appendix C. Proofs of the theorems

Proof of theorem 1

To prove the consistency of QMLE $\hat{\theta}$, it is sufficient to show that the following two facts hold (See white (1994)). The first one is the identification uniqueness condition: $\limsup_{n\to\infty}\Big[\max_{\theta\in B^c(\theta_0,\varepsilon)\cap\Theta} E\log L(\theta) - E\log L(\theta_0)\Big] < 0$ for any $\varepsilon > 0$, where $B^c(\theta_0,\varepsilon)$ is the compliment of an ε-neighborhood of θ_0. The second one is the uniform convergence in probability: $\sup_{\theta\in\Theta}\Big|\frac{1}{n}\log L(\theta) - \frac{1}{n}E\log L(\theta)\Big| = o_p(1)$.

The identification uniqueness

Firstly, we will show that the identification uniqueness condition hold. From the definition of the concentrated expected log-likelihood function, we have

$$\begin{aligned}\frac{1}{n}(E\log L(\theta) - E\log L(\theta_0)) &= -\frac{1}{2}\log(\tilde{\sigma}_0^2(\theta)) - \frac{1}{2n}\log|\Omega(\theta)| + \frac{1}{2}\log(\sigma_{00}^2(\theta)) + \frac{1}{2n}\log|\Omega(\theta_0)|,\\ &= -\frac{1}{2n}\log|\tilde{\sigma}_0^2(\theta)I_n| - \frac{1}{2n}\log|\Omega(\theta)| + \frac{1}{2n}\log|\sigma_{00}^2(\theta)I_n| + \frac{1}{2n}\log|\Omega(\theta_0)|,\\ &= \frac{1}{2n}\log|\sigma_{00}^2\Omega(\theta_0)| - \frac{1}{2n}\log|\hat{\sigma}_0^2(\theta)\Omega(\theta)|.\end{aligned}$$

By Assumption 7, for any $\theta \neq \theta_0$,

$$\lim_{n\to\infty}\frac{1}{n}\Big[\log|\sigma_{00}^2\Omega(\theta_0)| - \log|\tilde{\sigma}_0^2(\theta)\Omega(\theta)|\Big] \neq 0.$$

Thus,

$$\lim_{n\to\infty}\frac{1}{n}(E\log L(\theta) - E\log L(\theta_0)) \neq 0.$$

Let $p_n(\beta,\sigma_0^2,\theta) = \exp(\log L(\beta,\sigma_0^2,\theta))$ be the quasi-joint p.d.f of $u_0 = (Y - X\beta_0)$ and $p_n^0(\beta,\sigma_0^2,\theta)$ be the true joint p.d.f. We denote E^q as the expectation with respect to $p_n(\beta,\sigma_0^2,\theta)$ and E as the expectation with respect to $p_n^0(\beta,\sigma_0^2,\theta)$.

By the Jensen's inequality,

$$0 = \log E^q\left(\frac{p_n(\beta,\sigma_0^2,\theta)}{p_n(\beta_0,\sigma_{00}^2,\theta_0)}\right) \geq E^q\log\left(\frac{p_n(\beta,\sigma_0^2,\theta)}{p_n(\beta_0,\sigma_{00}^2,\theta_0)}\right)$$

Here, we note that u_0 appears in a quadratic form or linear-quadratic form in $p_n(\beta_0,\sigma_{00}^2,\theta_0)$ and $p_n(\beta,\sigma_0^2,\theta)$.

Thus,

$$E^q \log\left(\frac{p_n(\beta,\sigma_0^2,\theta)}{p_n(\beta_0,\sigma_{00}^2,\theta_0)}\right) = E\log\left(\frac{p_n(\beta,\sigma_0^2,\theta)}{p_n(\beta_0,\sigma_{00}^2,\theta_0)}\right).$$

This implies

$$E\log L(\theta) = \max_{\beta,\sigma_0^2} E[\log L(\beta,\sigma_0^2,\theta)] \le E[\log L(\beta_0,\sigma_{00}^2,\theta_0)] = E\log L(\theta_0).$$

Collecting the above results, we have

$$\lim_{n\to\infty}\left[\max_{\theta\in B^c(\theta_0,\varepsilon)\cap\Theta} E\log L(\theta) - E\log L(\theta_0)\right] < 0,$$

for any $\varepsilon > 0$, where $B^c(\theta_0,\varepsilon)$ is the compliment of an ε-neighborhood of θ_0. The identification uniqueness condition holds.

Uniform convergence

Secondly, we will show that the uniform convergence condition hold. From the definition, we have

$$\frac{1}{n}\log L(\theta) - \frac{1}{n}E\log L(\theta) = -\frac{1}{2}\log\hat{\sigma}_0^2 + \frac{1}{2}\log\tilde{\sigma}_0^2.$$

By the mean value theorem,

$$|\log\hat{\sigma}_0^2 - \log\tilde{\sigma}_0^2| = \frac{1}{\bar{\sigma}_0^2}|\hat{\sigma}_0^2 - \tilde{\sigma}_0^2|,$$

where $\bar{\sigma}_0^2$ lies between $\hat{\sigma}_0^2$ and $\tilde{\sigma}_0^2$. It is sufficient to show the following two facts. The first one is $\tilde{\sigma}_0^2$ is uniformly bounded away from zero and the second one is uniform convergence of $|\hat{\sigma}_0^2 - \tilde{\sigma}_0^2|$ in probability.

Firstly, we will show that $\tilde{\sigma}_0^2$ is uniformly bounded away from zero. By Assumption 4,

$$\begin{aligned}
\inf_{\theta\in\Theta}\tilde{\sigma}_0^2(\theta) &= \inf_{\theta\in\Theta}\left(\frac{\sigma_{00}^2}{n}tr(\Omega(\theta)^{-1}\Omega(\theta_0))\right),\\
&\ge \sigma_{00}^2\inf_{\theta\in\Theta}(\gamma_{\min}(\Omega^{-1}(\theta)))\frac{1}{n}tr(\Omega(\theta_0)),\\
&\ge c_0\underline{c}_\omega c_1,\\
&> 0,
\end{aligned}$$

where c_0 and c_1 are some constants. Therefore, $\tilde{\sigma}_0^2$ must be uniformly bounded away from zero.

Secondly, we will show that $\sup_{\theta\in\Theta} |\hat{\sigma}_0^2 - \tilde{\sigma}_0^2| = o_p(1)$. Because $M(\theta)\Omega^{-\frac{1}{2}}(\theta)X = 0$,

$$\begin{aligned}
\hat{\sigma}_0^2 - \tilde{\sigma}_0^2 &= \frac{1}{n}Y'\Omega^{-\frac{1}{2}}(\theta)M(\theta)\Omega^{-\frac{1}{2}}(\theta)Y - \frac{1}{n}E[u_0'\Omega^{-\frac{1}{2}}(\theta)M(\theta)\Omega^{-\frac{1}{2}}(\theta)u_0'] - \frac{1}{n}E[u_0'\Omega^{-\frac{1}{2}}(\theta)P(\theta)\Omega^{-\frac{1}{2}}(\theta)u_0'], \\
&= \frac{1}{n}\bigg(u_0'\Omega^{-\frac{1}{2}}(\theta)M(\theta)\Omega^{-\frac{1}{2}}(\theta)u_0 - Eu_0'\Omega^{-\frac{1}{2}}(\theta)M(\theta)\Omega^{-\frac{1}{2}}(\theta)u_0\bigg) - \frac{1}{n}E\bigg(u_0'\Omega^{-\frac{1}{2}}(\theta)P(\theta)\Omega^{-\frac{1}{2}}(\theta)u_0\bigg).
\end{aligned}$$

We will consider the uniform convergence of the above two terms.

Let us consider the uniform convergence of the second term. We note that $0 < \underline{c}_\omega\underline{c}_x \leq \inf_{\theta\in\Theta}\gamma_{\min}(\Omega^{-1}(\theta))\gamma_{\min}\left(\frac{X'X}{n}\right) \leq \gamma_{\min}\left(\frac{X'\Omega^{-1}(\theta)X}{n}\right) \leq \gamma_{\max}\left(\frac{X'\Omega^{-1}(\theta)X}{n}\right) \leq \sup_{\theta\in\Theta}\gamma_{\max}(\Omega^{-1}(\theta))\gamma_{\max}\left(\frac{X'X}{n}\right) \leq \bar{c}_\omega\bar{c}_x < \infty$. By assumption 4 and 5,

$$\begin{aligned}
\sup_{\theta\in\Theta}\left|\frac{1}{n}E\bigg(u_0'\Omega^{-\frac{1}{2}}(\theta)P(\theta)\Omega^{-\frac{1}{2}}(\theta)u_0\bigg)\right| &= \sup_{\theta\in\Theta}\left|\frac{1}{n}\sigma_{00}^2 tr\bigg(\Omega^{-1}(\theta)X(X'\Omega^{-1}(\theta)X)^{-1}X'\Omega^{-1}(\theta)\Omega(\theta_0)\bigg)\right|, \\
&\leq \frac{1}{n}\sup_{\theta\in\Theta}\left|\sigma_{00}^2\gamma_{\max}\left(\frac{X'\Omega^{-1}(\theta)X}{n}\right)^{-1}\gamma_{\max}(\Omega^{-2}(\theta))\right|\gamma_{\max}(\Omega(\theta_0))\frac{1}{n}tr(XX'), \\
&\leq \frac{1}{n}\sigma_{00}^2\sup_{\theta\in\Theta}\left|\gamma_{\max}\left(\frac{X'\Omega^{-1}(\theta)X}{n}\right)^{-1}\right| \\
&\sup_{\theta\in\Theta}\left|\gamma_{\max}(\Omega^{-2}(\theta))\right|\gamma_{\max}(\Omega(\theta_0))\frac{1}{n}tr(XX'), \\
&= \frac{1}{n}O(1)O(1)O(1)O(1)O(1), \\
&= o(1).
\end{aligned}$$

This implies that the second term converges uniformly.

To show that the uniform convergence of the first term, we will show that the pointwise convergence and stochastic equicontinuity of the term (See Andrew (1992)).

Firstly, we will consider the pointwise convergence of the first term. From Assumption 4 and 5, $\Omega^{-1}(\theta)$ and $X(X'\Omega^{-1}(\theta)X)^{-1}X'$ are uniformly bounded in both row and column sums. Therefore, $\Omega^{-\frac{1}{2}}(\theta)M(\theta)\Omega^{-\frac{1}{2}}(\theta) = \Omega^{-1}(\theta) - \Omega^{-1}(\theta)X(X'\Omega^{-1}(\theta)X)^{-1}X'\Omega^{-1}(\theta)$ is uniformly bounded in both row and column sums. By Lemma 3, it follows hat $\frac{1}{n}\bigg(u_0'\Omega^{-\frac{1}{2}}(\theta)M(\theta)\Omega^{-\frac{1}{2}}(\theta)u_0 - Eu_0'\Omega^{-\frac{1}{2}}(\theta)M(\theta)\Omega^{-\frac{1}{2}}(\theta)u_0\bigg) = o_p(1)$. This implies the first term converges pointwise.

Next, we will consider the stochastic equicontinuity condition of the first term. By the mean value theorem,

$$
\begin{aligned}
&\frac{1}{n}u_0'\Omega^{-\frac{1}{2}}(\theta_1)M(\theta_1)\Omega^{-\frac{1}{2}}(\theta_1)u_0 - \frac{1}{n}u_0'\Omega^{-\frac{1}{2}}(\theta_2)M(\theta_2)\Omega^{-\frac{1}{2}}(\theta_2)u_0 \\
&= \frac{1}{n}\sum_{i=1}^{p}\frac{\partial u_0'\Omega^{-\frac{1}{2}}(\bar{\theta})M(\bar{\theta})\Omega^{-\frac{1}{2}}(\bar{\theta})u_0}{\partial \rho_i}(\rho_{i,1}-\rho_{i,2}) + \frac{1}{n}\sum_{i=1}^{p}\frac{\partial u_0'\Omega^{-\frac{1}{2}}(\bar{\theta})M(\bar{\theta})\Omega^{-\frac{1}{2}}(\bar{\theta})u_0}{\partial \tau_i^2}(\tau_{i,1}^2-\tau_{i,2}^2),
\end{aligned}
$$

where $\bar{\theta}$ lies between θ_1 and θ_2. Thus, it is suffice to show that $\sup_{\theta\in\Theta}\left|\frac{1}{n}\frac{\partial u_0'\Omega^{-\frac{1}{2}}(\theta)M(\theta)\Omega^{-\frac{1}{2}}(\theta)u_0}{\partial \rho_i}\right| = O_p(1)$ and $\sup_{\theta\in\Theta}\left|\frac{1}{n}\frac{\partial u_0'\Omega^{-\frac{1}{2}}(\theta)M(\theta)\Omega^{-\frac{1}{2}}(\theta)u_0}{\partial \tau_i}\right| = O_p(1)$ (See, Davidoson (1994)).

Here, we note that the partial derivatives of $\Omega^{-1}(\theta)$ are given by,

$$
\begin{aligned}
\frac{\partial \Omega^{-1}(\theta)}{\partial \rho_i} &= -\tau_i^2\Omega^{-1}(\theta)JU_iA_i^{-1}(\rho_i)B_i(\rho_i)A_i^{-1}(\rho_i)U_i'J'\Omega^{-1}(\theta), \\
\frac{\partial \Omega^{-1}(\theta)}{\partial \tau_i} &= -\Omega^{-1}(\theta)JU_iA_i^{-1}(\rho_i)U_i'J'\Omega^{-1}(\theta),
\end{aligned}
$$

where $B_i(\rho_i) = W_i'(I_i-\rho_iW_i)+(I_i-\rho_iW_i')W_i$.

Let us consider the uniform boundedness of $\frac{1}{n}\frac{\partial u_0'\Omega^{-\frac{1}{2}}(\theta)M(\theta)\Omega^{-\frac{1}{2}}(\theta)u_0}{\partial \rho_i}$. The matrix $\Omega^{-\frac{1}{2}}(\theta)M(\theta)\Omega^{-\frac{1}{2}}(\theta)$ consists of the two termes $\Omega^{-1}(\theta)$ and $\Omega^{-1}(\theta)X(X'\Omega^{-1}(\theta)X)^{-1}X'\Omega^{-1}(\theta)$. The uniform boundedness of $\frac{1}{n}\frac{\partial u_0'\Omega^{-1}(\theta)u_0}{\partial \rho_i}$ is given by,

$$
\begin{aligned}
\sup_{\theta\in\Theta}\left|\frac{1}{n}\frac{\partial u_0'\Omega^{-1}(\theta)u_0}{\partial \rho_i}\right| &= \sup_{\theta\in\Theta}\left|\frac{1}{n}\tau_i^2u_0'\Omega^{-1}(\theta)JU_iA_i^{-1}(\rho_i)B_i(\rho_i)A_i^{-1}(\rho_i)U_i'J'\Omega^{-1}(\theta)u_0\right|, \\
&= \sup_{\theta\in\Theta}\left|\tau_i^2\gamma_{\max}(JU_iA_i^{-1}(\rho_i)B_i^{-1}(\rho_i)A_i^{-1}(\rho_i)U_i'J')\gamma_{\max}^2(\Omega^{-1}(\theta))\right|\frac{1}{n}u_0'u_0, \\
&= O(1)O(1)O(1)O_p(1), \\
&= O_p(1).
\end{aligned}
$$

Next, we will show that the uniform boundness of $\frac{1}{n}\frac{\partial u_0'\Omega^{-1}(\theta)X(X'\Omega^{-1}(\theta)X)^{-1}X'\Omega^{-1}(\theta)u_0}{\partial \rho_i}$. The partial derivative of the matrix is

$$
\begin{aligned}
\frac{1}{n}\frac{\partial u_0'\Omega^{-1}(\theta)X(X'\Omega^{-1}(\theta)X)^{-1}X'\Omega^{-1}(\theta)u_0}{\partial \rho_i} &= \frac{1}{n}u_0'\frac{\partial \Omega^{-1}(\theta)}{\partial \rho_i}X(X'\Omega^{-1}(\theta)X)^{-1}X'\Omega^{-1}(\theta)u_0 \\
&+ \frac{1}{n}u_0'\Omega^{-1}(\theta)X\frac{\partial (X'\Omega^{-1}(\theta)X)^{-1}}{\partial \rho_i}X'\Omega^{-1}(\theta)u_0 \\
&+ \frac{1}{n}u_0'\Omega^{-1}(\theta)X(X'\Omega^{-1}(\theta)X)^{-1}X'\frac{\partial \Omega^{-1}(\theta)}{\partial \rho_i}u_0, \\
&= \phi_1+\phi_2+\phi_3,
\end{aligned}
$$

where $\phi_1 = \frac{1}{n}u_0'\frac{\partial\Omega^{-1}(\theta)}{\partial\rho_i}X(X'\Omega^{-1}(\theta)X)^{-1}X'\Omega^{-1}(\theta)u_0, \phi_2 = \frac{1}{n}u_0'\Omega^{-1}(\theta)X(X'\Omega^{-1}(\theta)X)^{-1}X'\frac{\partial\Omega^{-1}(\theta)}{\partial\rho_i}u_0$ and $\phi_3 = \frac{1}{n}u_0'\Omega^{-1}(\theta)X(X'\Omega^{-1}(\theta)X)^{-1}X'\frac{\partial\Omega^{-1}(\theta)}{\partial\rho_i}u_0$.

By Lemma 3, the uniform boundness of ϕ_2 is given by

$$
\begin{aligned}
&\sup_{\theta\in\Theta}\left|\frac{1}{n}u_0'\Omega^{-1}(\theta)X\frac{\partial(X'\Omega^{-1}(\theta)X)^{-1}}{\partial\rho_i}X'\Omega^{-1}(\theta)u_0\right|,\\
&=\sup_{\theta\in\Theta}\left|\frac{1}{n}\tau_i^2u_0'\Omega^{-1}(\theta)X(X'\Omega^{-1}(\theta)X)^{-1}X'\Omega^{-1}(\theta)JU_iA_i^{-1}(\rho_i)B_i(\rho_i)A_i^{-1}(\rho_i)U_i'J'\Omega^{-1}(\theta)X(X'\Omega^{-1}(\theta)X)^{-1}X'\Omega^{-1}(\theta)u_0\right|,\\
&\le\sup_{\theta\in\Theta}\left|\tau_i^2\gamma_{\max}(JU_iA_i^{-1}(\rho_i)B_i(\rho_i)A_i^{-1}(\rho_i)U_i'J')\gamma_{\max}(\Omega^{-1}(\theta))\gamma_{\max}(\frac{(X'\Omega^{-1}(\theta)X)^{-1}}{n})\gamma_{\max}(\frac{X'X}{n})\gamma_{\max}(\Omega^{-2}(\theta))\right|\frac{1}{n}u_0'u_0,\\
&=O(1)O(1)O(1)O(1)O(1)O(1)O_p(1),\\
&=O_p(1).
\end{aligned}
$$

Let us consider the uniform boundness of ϕ_1. The term is

$$
\begin{aligned}
&\frac{1}{n}u_0'\frac{\partial\Omega^{-1}(\theta)}{\partial\rho_i}X(X'\Omega^{-1}(\theta)X)^{-1}X'\Omega^{-1}(\theta)u_0,\\
&=\tau_i^2\frac{1}{n}u_0'\Omega^{-1}(\theta)JU_iA_i^{-1}(\rho_i)B_i(\rho_i)A_i^{-1}(\rho_i)U_i'J'\Omega^{-1}(\theta)X(X'\Omega^{-1}(\theta)X)^{-1}X'\Omega^{-1}(\theta)u_0,\\
&=tr(a'(\theta)b(\theta)),
\end{aligned}
$$

where $a'(\theta) = \tau_i^2\frac{1}{\sqrt{n}}u_0'\Omega^{-1}(\theta)JU_iA_i^{-1}(\rho_i)B_i(\rho_i)A_i^{-1}(\rho_i)U_i'J'$ and $b(\theta) = \frac{1}{\sqrt{n}}\Omega^{-1}(\theta)X(X'\Omega^{-1}(\theta)X)^{-1}X'\Omega^{-1}(\theta)u_0$ It suffices to show that $\left(\sup_{\theta\in\Theta}\left|tr(a'(\theta)b(\theta))\right|\right)^2 = O_p(1)$. Because of $(\sup_{\theta\in\Theta}f(\theta))^2 = \sup_{\theta\in\Theta}f(\theta)^2$, by Cauchy-Schwarz inequality,

$$
\begin{aligned}
\left(\sup_{\theta\in\Theta}\left|tr(a'(\theta)b(\theta))\right|\right)^2 &= \sup_{\theta\in\Theta}tr^2(a'(\theta)b(\theta)),\\
&\le\sup_{\theta\in\Theta}tr(a'(\theta)a(\theta))\sup_{\theta\in\Theta}tr(b'(\theta)b(\theta)).
\end{aligned}
$$

The uniform boundenss of $tr(a'(\theta)a(\theta))$ is given by,

$$
\begin{aligned}
&\sup_{\theta\in\Theta} tr(a'(\theta)a(\theta)) \\
&= \sup_{\theta\in\Theta}\left(\tau_i^2\frac{1}{n}tr(u_0'\Omega^{-1}(\theta)JU_iA_i^{-1}(\rho_i)B_i(\rho_i)A_i^{-1}(\rho_i)U_i'J'JU_iA_i^{-1}(\rho_i)B_i(\rho_i)A_i^{-1}(\rho_i)U_i'J'\Omega^{-1}(\theta)u_0)\right), \\
&\leq \sup_{\theta\in\Theta}\left(\tau_i^2\gamma_{\max}^2(JU_iA_i^{-1}(\rho_i)B_i(\rho_i)A_i^{-1}(\rho_i)U_i'J')\gamma_{\max}^2(\Omega^{-1}(\theta))\right)\frac{1}{n}u_0'u_0, \\
&= O(1)O(1)O(1)O_p(1), \\
&= O_p(1).
\end{aligned}
$$

Similarly, the uniform boundness of $tr(b'(\theta)b(\theta))$ is given by,

$$
\begin{aligned}
&\sup_{\theta\in\Theta} tr(b'(\theta)b(\theta)) \\
&= \sup_{\theta\in\Theta}\left(\frac{1}{n}tr(u_0'\Omega^{-1}(\theta)X(X'\Omega^{-1}(\theta)X)^{-1}X'\Omega^{-1}(\theta)\Omega^{-1}(\theta)X(X'\Omega^{-1}(\theta)X)^{-1}X'\Omega^{-1}(\theta)u_0)\right), \\
&\leq \sup_{\theta\in\Theta}\left(\gamma_{\max}(\Omega^{-1}(\theta))\gamma_{\max}(\frac{(X'\Omega^{-1}(\theta)X)^{-1}}{n})\gamma_{\max}(\frac{X'X}{n})\gamma_{\max}^2(\Omega^{-1}(\theta))\right)\frac{1}{n}u_0'u_0, \\
&= O(1)O(1)O(1)O(1)O_p(1), \\
&= O_p(1).
\end{aligned}
$$

The uniform boundedness of $\frac{1}{n}\frac{\partial u_0'\Omega^{-\frac{1}{2}}(\theta)M(\theta)\Omega^{-\frac{1}{2}}(\theta)u_0}{\partial\tau_i}$ can be proved by the similar manner. By collecting above results, we can show that the uniform convergence of $\frac{1}{n}\left(u_0'\Omega^{-\frac{1}{2}}(\theta)M(\theta)\Omega^{-\frac{1}{2}}(\theta)u_0 - Eu_0'\Omega^{-\frac{1}{2}}(\theta)M(\theta)\Omega^{-\frac{1}{2}}(\theta)u_0\right)$. Therefore, $\sup_{\theta\in\Theta}\left|\frac{1}{n}\log L(\theta) - \frac{1}{n}E\log L(\theta)\right| = o_p(1)$, and the QMLE $\hat{\theta}$ is a consistent estimator of θ_0 by White (1994).

Proof of Theorem 2

To derive the asymptotic normality of the proposed estimator, we will show that the following three results:

1. $\frac{1}{n}\frac{\partial^2\log L(\psi_0)}{\partial\psi\partial\psi'} \xrightarrow{p} E\left[\frac{1}{n}\frac{\partial^2\log L(\psi_0)}{\partial\psi\partial\psi'}\right]$.
2. $\frac{1}{n}\frac{\partial^2\log L(\bar{\psi})}{\partial\psi\partial\psi'} \xrightarrow{p} \frac{1}{n}\frac{\partial^2\log L(\psi_0)}{\partial\psi\partial\psi'}$.
3. $\frac{1}{\sqrt{n}}\frac{\partial\log L(\psi_0)}{\partial\psi} \xrightarrow{D} N(0,\Gamma)$.

If we can hold above results, then the asymptotic distribution of $\hat{\psi}$ the following distribution. By the mean value theorem,

$$\sqrt{n}(\hat{\psi} - \psi_0) = -\left(\frac{1}{n}\frac{\partial^2 \log L(\bar{\psi})}{\partial\psi\partial\psi'}\right)^{-1} \frac{1}{\sqrt{n}}\frac{\partial \log L(\psi_0)}{\partial\psi},$$

where $\bar{\psi}$ lies between $\hat{\psi}$ and ψ_0. Let $\Sigma = -\lim_{n\to\infty} E\left(\frac{1}{n}\frac{\partial^2 \log L(\psi_0)}{\partial\psi\partial\psi'}\right)$ and $\Gamma^* = \lim_{n\to\infty}\frac{1}{n}\Gamma$. Thus, we obtain,

$$\sqrt{n}(\hat{\psi} - \psi_0) \xrightarrow{D} N(0, \Sigma^{-1}\Gamma^*\Sigma^{-1}).$$

Proof of result1: $\frac{1}{n}\frac{\partial^2 \log L(\psi_0)}{\partial\psi\partial\psi'} \xrightarrow{p} E\left[\frac{1}{n}\frac{\partial^2 \log L(\psi_0)}{\partial\psi\partial\psi'}\right]$

Let us consider $\frac{1}{n}\frac{\partial^2 \log L(\psi_0)}{\partial\psi\partial\psi'} \xrightarrow{p} E\left[\frac{1}{n}\frac{\partial^2 \log L(\psi_0)}{\partial\psi\partial\psi'}\right]$. The terms of the hessian matrix is decomposed into the forms:

1. $\frac{1}{n}X'AX$,
2. $\frac{1}{n}tr(A)$,
3. $\frac{1}{n}X'Au_0$,
4. $\frac{1}{n}u_0'Au_0$,

where A is an $n \times n$ matrix whose row and column sums are uniformly bounded.

The convergences in probability of the first two terms are clear. The third term is given by,

$$\frac{1}{n}X'Au_0 = \frac{1}{n}X'A\varepsilon + \sum_{k=1}^{p}\frac{1}{n}X'AJU_k(I_k - \rho_k W_k)^{-1}f_k.$$

The row and column sums of $AJU_k(I_k - \rho_k W_k)^{-1}$ is uniformly bounded. From Lemma 1,

$$\begin{aligned}\frac{1}{n}X'A\varepsilon &= \frac{1}{\sqrt{n}}O_p(1) = o_p(1),\\ \frac{1}{n}X'AJU_k(I_k - \rho_k W_k)^{-1}f_k &= \frac{m_k}{n}\frac{1}{\sqrt{m_k}}O_p(1) = o_p(1).\end{aligned}$$

Thus, $\frac{1}{n}X'Au_0 - E(\frac{1}{n}X'Au_0) = o_p(1)$. It follows that $\frac{1}{n}u_0'Au_0 - E\left[\frac{1}{n}u_0'Au_0\right]$ by Lemma 3. Collecting above results, we obtain $\frac{1}{n}\frac{\partial^2 \log L(\psi_0)}{\partial\psi\partial\psi'} \xrightarrow{p} E\left[\frac{1}{n}\frac{\partial^2 \log L(\psi_0)}{\partial\psi\partial\psi'}\right]$.

Proof of result 2: $\frac{1}{n}\frac{\partial^2 \log L(\bar{\psi})}{\partial\psi\partial\psi'} \xrightarrow{p} \frac{1}{n}\frac{\partial^2 \log L(\psi_0)}{\partial\psi\partial\psi'}$

We will show that $\frac{1}{n}\frac{\partial^2 \log L(\bar{\psi})}{\partial\psi\partial\psi'} \xrightarrow{p} \frac{1}{n}\frac{\partial^2 \log L(\psi_0)}{\partial\psi\partial\psi'}$. Firstly, let us consider the case of $\frac{\partial^2 \log L(\psi)}{\partial\beta\partial\beta'}$. By the mean value theorem,

$$\begin{aligned}X'\Sigma^{-1}(\bar{\eta})X = X'\Sigma^{-1}(\eta_0)X &+ \sum_{k=0}^{p} X'\frac{\partial\Sigma^{-1}(\tilde{\eta})}{\partial\sigma_k^2}X(\sigma_{0k}^2 - \bar{\sigma}_k^2)\\ &+ \sum_{k=1}^{p} X'\frac{\partial\Sigma^{-1}(\tilde{\eta})}{\partial\rho_k}X(\rho_{0k} - \bar{\rho}_k),\end{aligned}$$

where $\tilde{\eta}$ lies between η_0 and $\bar{\eta}$. Because the element of $\frac{1}{n}X'\frac{\partial\Sigma^{-1}(\tilde{\eta})}{\partial\sigma_k^2}X$ is uniformly bounded,

$$\begin{aligned}\frac{1}{n}\frac{\partial^2 \log L(\bar{\psi})}{\partial\beta\partial\beta'} - \frac{1}{n}\frac{\partial^2 \log L(\psi_0)}{\partial\beta\partial\beta'} &= \sum_{k=0}^{p}\frac{1}{n}X'\frac{\partial\Sigma^{-1}(\tilde{\eta})}{\partial\sigma_k^2}X(\sigma_{0k}^2 - \bar{\sigma}_k^2) + \sum_{k=1}^{p}\frac{1}{n}X'\frac{\partial\Sigma^{-1}(\tilde{\eta})}{\partial\rho_k}X(\rho_{0k} - \bar{\rho}_k),\\ &= \sum_{k=0}^{p} O(1)o_p(1) + \sum_{k=1}^{p} O(1)o_p(1),\\ &= o_p(1).\end{aligned}$$

Secondly, we will consider the case of $\frac{\partial^2 \log L(\psi)}{\partial\beta\partial\sigma_i^2}$. By Lemma 3,

$$\begin{aligned}\frac{1}{n}\frac{\partial^2 \log L(\bar{\psi})}{\partial\beta\partial\sigma_i^2} - \frac{1}{n}\frac{\partial^2 \log L(\psi_0)}{\partial\beta\partial\sigma_i^2} &= \frac{1}{n}X'\Sigma^{-1}(\bar{\eta})G_i(\bar{\rho}_i)\Sigma^{-1}(\bar{\eta})(Y - X\bar{\beta}) - \frac{1}{n}X'\Sigma^{-1}(\eta_0)G_i(\rho_{0i})\Sigma^{-1}(\eta_0)(Y - X\beta_0),\\ &= \frac{1}{n}X'\Sigma^{-1}(\bar{\eta})G_i(\bar{\rho}_i)\Sigma^{-1}(\bar{\eta})X(\beta_0 - \bar{\beta})\\ &+ \frac{1}{n}X'\Sigma^{-1}(\bar{\eta})G_i(\bar{\rho}_i)\Sigma^{-1}(\bar{\eta})u_0 - \frac{1}{n}X'\Sigma^{-1}(\eta_0)G_i(\rho_{0i})\Sigma^{-1}(\eta_0)u_0,\\ &= O_p(1)o_p(1) + o_p(1) + o_p(1),\\ &= o_p(1).\end{aligned}$$

Next, let us consider the case of $\frac{\partial^2 \log L(\psi)}{\partial\sigma_i^2\partial\sigma_j^2}$. By the mean value theorem,

$$\begin{aligned}tr(\Sigma^{-1}(\bar{\eta})G_j(\bar{\rho}_j)\Sigma^{-1}(\bar{\eta})G_i(\bar{\rho})) = tr(\Sigma^{-1}(\eta_0)G_j(\rho_{0j})\Sigma^{-1}(\eta_0)G_i(\rho_{0i}))&\\ + \sum_{k=0}^{p}\frac{\partial tr(\Sigma^{-1}(\tilde{\eta})G_j(\tilde{\rho}_j)\Sigma^{-1}(\tilde{\eta})G_i(\tilde{\rho}_i))}{\partial\sigma_k^2}(\sigma_{0k}^2 - \bar{\sigma}_k^2)&\\ + \sum_{k=1}^{p}\frac{\partial tr(\Sigma^{-1}(\tilde{\eta})G_j(\tilde{\rho}_j)\Sigma^{-1}(\tilde{\eta})G_i(\tilde{\rho}_i))}{\partial\rho_k}(\rho_{0k} - \bar{\rho}_k),&\end{aligned}$$

where $\tilde{\eta}$ lies between η_0 and $\bar{\eta}$. Because $\frac{\partial \Sigma^{-1}(\tilde{\eta})G_j(\tilde{\rho}_j)\Sigma^{-1}(\tilde{\eta})G_i(\tilde{\rho}_i))}{\partial \sigma_k^2}$ and $\frac{\partial \Sigma^{-1}(\tilde{\eta})G_j(\tilde{\rho}_j)\Sigma^{-1}(\tilde{\eta})G_i(\tilde{\rho}_i))}{\partial \rho_k}$ is uniformly bounded in both row and column sums, $\frac{\partial tr(\Sigma^{-1}(\tilde{\eta})G_j(\tilde{\rho}_j)\Sigma^{-1}(\tilde{\eta})G_i(\tilde{\rho}_i))}{\partial \sigma_k^2}$ and $\frac{\partial tr(\Sigma^{-1}(\tilde{\eta})G_j(\tilde{\rho}_j)\Sigma^{-1}(\tilde{\eta})G_i(\tilde{\rho}_i))}{\partial \rho_k}$ are $O(n)$ by Lee (2004). Thus,

$$
\begin{aligned}
&\frac{1}{2n}tr(\Sigma^{-1}(\bar{\eta})G_j(\bar{\rho}_j)\Sigma^{-1}(\bar{\eta})G_i(\bar{\rho}_i)) - \frac{1}{2n}tr(\Sigma^{-1}(\eta_0)G_j(\rho_{0j})\Sigma^{-1}(\eta_0)G_i(\rho_{0i}))\\
&= \frac{1}{2n}\sum_{k=0}^{p}\frac{\partial tr(\Sigma^{-1}(\tilde{\eta})G_j(\tilde{\rho}_j)\Sigma^{-1}(\tilde{\eta})G_i(\tilde{\rho}_i))}{\partial \sigma_k^2}(\sigma_{0k}^2 - \bar{\sigma}_k^2) + \frac{1}{2n}\sum_{k=1}^{p}\frac{\partial tr(\Sigma^{-1}(\tilde{\eta})G_j(\tilde{\rho}_j)\Sigma^{-1}(\tilde{\eta})G_i(\tilde{\rho}_i))}{\partial \rho_k}(\rho_{0k} - \bar{\rho}_k),\\
&= \frac{1}{2}\sum_{k=0}^{p}O(1)o_p(1) + \frac{1}{2}\sum_{k=1}^{p}O(1)o_p(1),\\
&= o_p(1).
\end{aligned}
$$

Similarly, by the mean value theorem,

$$
\begin{aligned}
\Sigma^{-1}(\bar{\eta})G_j(\bar{\rho}_j)\Sigma^{-1}(\bar{\eta})G_i(\bar{\rho}_i)\Sigma^{-1}(\bar{\eta}) &= \Sigma^{-1}(\eta_0)G_j(\rho_{0j})\Sigma^{-1}(\eta_0)G_i(\rho_{0i})\Sigma^{-1}(\eta_0)\\
&+ \sum_{k=0}^{p}\frac{\partial \Sigma^{-1}(\tilde{\eta})G_j(\tilde{\rho}_j)\Sigma^{-1}(\tilde{\eta})G_i(\tilde{\rho}_i)\Sigma^{-1}(\tilde{\eta})}{\partial \sigma_k^2}(\sigma_{0k}^2 - \bar{\sigma}_k^2)\\
&+ \sum_{k=1}^{p}\frac{\partial \Sigma^{-1}(\tilde{\eta})G_j(\tilde{\rho}_j)\Sigma^{-1}(\tilde{\eta})G_i(\tilde{\rho}_i)\Sigma^{-1}(\tilde{\eta})}{\partial \rho_k}(\rho_{0k} - \bar{\rho}_k)
\end{aligned}
$$

Because $\frac{\partial \Sigma^{-1}(\tilde{\eta})G_j(\tilde{\rho}_j)\Sigma^{-1}(\tilde{\eta})G_i(\tilde{\rho}_i)\Sigma^{-1}(\tilde{\eta})}{\partial \sigma_k^2}$ and $\frac{\partial \Sigma^{-1}(\tilde{\eta})G_j(\tilde{\rho}_j)\Sigma^{-1}(\tilde{\eta})G_i(\tilde{\rho}_i)\Sigma^{-1}(\tilde{\eta})}{\partial \rho_k}$ are uniformly bounded in both row and column sums, by Lemma 3,

$$
\begin{aligned}
&\frac{1}{n}(Y - X\bar{\beta})'\Sigma^{-1}(\bar{\eta})G_j(\bar{\rho}_j)\Sigma^{-1}(\bar{\eta})G_i(\bar{\rho}_i)\Sigma^{-1}(\bar{\eta})(Y - X\bar{\beta})\\
&- \frac{1}{n}(Y - X\beta_0)'\Sigma^{-1}(\eta_0)G_j(\rho_{0j})\Sigma^{-1}(\eta_0)G_i(\rho_{0i})\Sigma^{-1}(\eta_0)(Y - X\beta_0)\\
&= \frac{1}{n}(\beta_0 - \bar{\beta})'X'\Sigma^{-1}(\bar{\eta})G_j(\bar{\rho}_j)\Sigma^{-1}(\bar{\eta})G_i(\bar{\rho}_i)\Sigma^{-1}(\bar{\eta})X(\beta_0 - \bar{\beta}) + \frac{1}{n}(\beta_0 - \bar{\beta})'X'\Sigma^{-1}(\bar{\eta})G_j(\bar{\rho}_j)\Sigma^{-1}(\bar{\eta})G_i(\bar{\rho}_i)\Sigma^{-1}(\bar{\eta})u_0\\
&+ \frac{1}{n}u_0'\Sigma^{-1}(\bar{\eta})G_j(\bar{\rho}_j)\Sigma^{-1}(\bar{\eta})G_i(\bar{\rho}_i)\Sigma^{-1}(\bar{\eta})X(\beta_0 - \bar{\beta}) + \frac{1}{n}\sum_{k=0}^{p}u_0'\frac{\partial \Sigma^{-1}(\tilde{\eta})G_j(\tilde{\rho}_j)\Sigma^{-1}(\tilde{\eta})G_i(\tilde{\rho}_i)\Sigma^{-1}(\tilde{\eta})}{\partial \sigma_k^2}u_0(\sigma_{0k}^2 - \bar{\sigma}_k^2)\\
&+ \frac{1}{n}\sum_{k=1}^{p}u_0'\frac{\partial \Sigma^{-1}(\tilde{\eta})G_j(\tilde{\rho}_j)\Sigma^{-1}(\tilde{\eta})G_i(\tilde{\rho}_i)\Sigma^{-1}(\tilde{\eta})}{\partial \rho_k}u_0(\rho_{0k} - \bar{\rho}_k),\\
&= o_p(1)O(1)o_p(1) + o_p(1)o_p(1) + o_p(1)o_p(1) + \sum_{k=0}^{p}O_p(1)o_p(1) + \sum_{k=1}^{p}O_p(1)o_p(1),\\
&= o_p(1).
\end{aligned}
$$

The convergences in probability of the other elements of the hessian matrix can be shown in similar manner. Thus, $\frac{1}{n}\frac{\partial^2 \log L(\bar{\psi})}{\partial\psi\partial\psi'} \xrightarrow{p} \frac{1}{n}\frac{\partial^2 \log L(\psi_0)}{\partial\psi\partial\psi'}$.

Result 3: $\frac{1}{\sqrt{n}}\frac{\partial \log L(\psi_0)}{\partial\psi} \xrightarrow{D} N(0,\Gamma)$

Finally, we will show that $\frac{1}{\sqrt{n}}\frac{\partial \log L(\psi_0)}{\partial\psi} \xrightarrow{D} N(0, \lim_{n\to\infty}\frac{1}{n}\Gamma)$, where Γ is the variance of the score function. We will apply the Cramer-Wold devise to derive the joint asymptotic normality. Let $c = (c'_\beta, c'_{\sigma_i^2}, c'_{\rho_j})'$ be a nonzero (k + (p + 1) + p) vector of constants. Here, $c'\frac{\partial \log L(\psi_0)}{\partial\psi}$ can be written as

$$\begin{aligned}
c'\frac{\partial \log L(\psi_0)}{\partial\psi} &= c'_\beta\frac{\partial \log L(\psi_0)}{\partial\beta} + \sum_{i=0}^{p} c_{\sigma_i^2}\frac{\partial \log L(\psi_0)}{\partial\sigma_i^2} + \sum_{j=1}^{p} c_{\rho_j}\frac{\partial \log L(\psi_0)}{\partial\rho_j} \\
&= c'_\beta X'\Sigma^{-1}(\eta_0)u_0 + u'_0\Sigma^{-1}(\eta_0)\Bigg(\sum_{i=0}^{p}\frac{c_{\sigma_i^2}}{2}G_i(\rho_{0i}) - \sum_{j=1}^{p}\frac{c_{\rho_j}\sigma_j^2}{2}H_i(\rho_{0j})\Bigg)\Sigma^{-1}(\eta_0)u_0 \\
&\quad - E\Bigg[u'_0\Sigma^{-1}(\eta_0)\Bigg(\sum_{i=0}^{p}\frac{c_{\sigma_i^2}}{2}G_i(\rho_{0i}) - \sum_{j=1}^{p}\frac{c_{\rho_j}\sigma_j^2}{2}H_i(\rho_{0j})\Bigg)\Sigma^{-1}(\eta_0)u_0\Bigg].
\end{aligned}$$

We denote $v = (f'_0, \ldots, f'_p)$, $U_0 = I_n$, $W_0 = I_n$, $\rho_{00} = 0$, $b_i = (I_i - \rho_i W_i)^{-1}U_i\Sigma^{-1}(\eta_0)Xc_\beta$, $b = (b'_0, \ldots, b'_p)'$, $A_{i,j} = (I_i - \rho_{0i}W'_i)^{-1}U'_i\Sigma^{-1}(\eta_0)\Bigg(\sum_{i=0}^{p}\frac{c_{\sigma_i^2}}{2}G_i(\rho_{0i}) - \sum_{j=1}^{p}\frac{c_{\rho_j}\sigma_j^2}{2}H_i(\rho_{0j})\Bigg)\Sigma^{-1}(\eta_0)U_j(I_j - \rho_{0j}W_j)^{-1}$ and

$$A = \begin{pmatrix} A_{00} & A_{01} & \ldots & A_{0p} \\ A_{10} & A_{11} & \ldots & A_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ A_{p0} & A_{p1} & \ldots & A_{pp} \end{pmatrix}.$$

Then, the linear combination of the elements of the score vector is given by the following linear-quadratic equation:

$$c'\frac{\partial \log L(\psi_0)}{\partial\psi} = b'v + v'Av - E(v'Av).$$

We denote b_i is the i-th element of b. Because the element of b is uniformly bounded, there exists a constantc_b such that $|b_i| \leq c_b$ for all i. It follow that $\frac{1}{n}\sum_{i=1}^{n}|b_i|^{2+\delta} \leq c_b^{2+\delta_1}$ for some $\delta_1 > 0$, and thus $\sup_n \frac{1}{n}\sum_{i=1}^{n}|b_i|^{2+\delta} < \infty$. Because the linear-quadratic form, $c'\frac{\partial \log L(\psi_0)}{\partial\psi}$, holds the assumption of Theorem 1 in Kelejian and Prucha (2001), $c'\frac{\partial \log L(\psi_0)}{\partial\psi} \xrightarrow{D} N(0, c'\Gamma c)$. By Cramer-Wold theorem, it follows that $\frac{1}{\sqrt{n}}\frac{\partial \log L(\psi_0)}{\partial\psi} \xrightarrow{D} N(0, \lim_{n\to\infty}\frac{1}{n}\Gamma)$.