Web Supplementary Material For
"Efficient Pairwise Composite Likelihood Estimation for Spatial-Clustered Data"

Yun Bai, Jian Kang and Peter X.-K. Song

1 Matérn Covariance Function

In this section, we define the Matérn covariance function used in the paper and provide relevant formulas for derivatives.

Definition The Matérn covariance function is defined as

$$C_\nu(d;\alpha) = \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{\sqrt{2\nu}d}{\alpha}\right)^\nu K_\nu\left(\sqrt{2\nu}\frac{d}{\alpha}\right), \tag{i}$$

where $K_\nu(x)$ is the modified Bessel function of the second kind, which can be expressed as

$$K_\nu(x) \;=\; \int_0^\infty \exp[-x\cosh(t)]\cosh(\nu t)dt, \tag{ii}$$

and $\cosh(t)$ is the hyperbolic cosine function: i.e. $\cosh(t) = (e^t + e^{-t})/2$.

Note that $C_\nu(0,\alpha) = 1$ and $C_{0.5}(d;\alpha) = \sqrt{\frac{2}{\pi}}\sqrt{\frac{d}{\alpha}}K_{0.5}\left(\frac{d}{\alpha}\right) = \exp\left(-\frac{d}{\alpha}\right)$, This is equivalent to the exponential decay covariance function.

Proposition 1.1 *The first order derivative of $K_\nu(x)$ satisfies with the following equations:*

$$\frac{dK_\nu(x)}{dx} = -\frac{1}{2}\left[K_{\nu+1}(x) + K_{\nu-1}(x)\right] \quad \text{for } \nu \geq 1, \tag{iii}$$

and $\frac{dK_0(x)}{dx} = -K_1(x)$.

Proof : By definition, the first derivative of the $K_\nu(x)$ is

$$\frac{dK_\nu(x)}{dx} = \int_0^\infty \frac{d\exp[-x\cosh(t)]}{dx}\cosh(\nu t)dt$$
$$= -\int_0^\infty \cosh(t)\exp(-x\cosh(t))\cosh(\nu t)dt$$

Note that $\cosh[(\nu \pm 1)t] = \cosh(t)\cosh(\nu t) \pm \sinh(t)\sinh(\nu t)$ and

$\cosh(t)\cosh(\nu t) = \frac{1}{2}\{\cosh[(\nu+1)t] + \cosh[(\nu-1)t]\}$. Thus,

$$\begin{aligned} K_\nu(x) &= \int_0^\infty \exp[-x\cosh(t)]\cosh(\nu t)dt \\ &= \frac{1}{\nu}\int_0^\infty \exp[-x\cosh(t)]d\sinh(\nu t) \\ &= \frac{1}{\nu}\left[\exp[-x\cosh(t)]\sinh(\nu t)\Big|_0^\infty - \int_0^\infty \sinh(\nu t)d\exp[-x\cosh(t)]\right] \\ &= \frac{x}{\nu}\int_0^\infty \sinh(t)\sinh(\nu t)\exp[-x\cosh(t)]dt \\ &= \frac{x}{2\nu}\left[K_{\nu+1}(x) - K_{\nu-1}(x)\right]. \end{aligned}$$

Furthermore,

$$K_{\nu-1}(x) = K_{\nu+1}(x) - \frac{2\nu}{x}K_\nu(x) \quad \text{and} \quad \frac{\mathrm{d}K_\nu(x)}{\mathrm{d}x} = \frac{\nu}{x}K_v(x) - K_{\nu+1}(x).$$

This completes the proof. ∎

Proposition 1.2 *The first order derivative of the Matérn covariance function with respect to α has the following representation:*

$$\frac{dC_\nu(d;\alpha)}{d\alpha} = \frac{2\nu}{\alpha}\left[C_{\nu+1}\left(\sqrt{\frac{\nu}{\nu+1}}d;\alpha\right) - C_\nu(d;\alpha)\right]. \tag{iv}$$

Proof :

$$\frac{dC_\nu(d;\alpha)}{d\alpha} = \frac{\mathrm{d}\left[\frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{\sqrt{2\nu}d}{\alpha}\right)^\nu K_\nu\left(\sqrt{2\nu}\frac{d}{\alpha}\right)\right\}}{\mathrm{d}\alpha} = \frac{(\sqrt{2\nu}d)^\nu}{\Gamma(\nu)2^{\nu-1}}\frac{\mathrm{d}\left[\alpha^{-\nu}K_\nu(\sqrt{2\nu}d\alpha^{-1})\right]}{\mathrm{d}\alpha}$$
$$= \frac{(\sqrt{2\nu}d)^\nu}{\Gamma(\nu)2^{\nu-1}}\left[-\nu\alpha^{-\nu-1}K_\nu(\sqrt{2\nu}d\alpha^{-1}) + \alpha^{-\nu}(-\sqrt{2\nu}d\alpha^{-2})K'_\nu(\sqrt{2\nu}d\alpha^{-1})\right]$$
$$= -\frac{(\sqrt{2\nu}d)^\nu}{\Gamma(\nu)2^{\nu-1}}\left[\nu\alpha^{-\nu-1}K_\nu(\sqrt{2\nu}d\alpha^{-1}) + \alpha^{-\nu-2}(\sqrt{2\nu}d)\frac{\mathrm{d}K_\nu(\sqrt{2\nu}d\alpha^{-1})}{\mathrm{d}\alpha}\right]$$

By equation (iii), we have

$$\frac{\mathrm{d}K_\nu(\sqrt{2\nu}d\alpha^{-1})}{\mathrm{d}\alpha} = \left[\frac{\nu\alpha}{\sqrt{2\nu}d}K_\nu(\sqrt{2\nu}d\alpha^{-1}) - K_{\nu+1}(\sqrt{2\nu}d\alpha^{-1})\right].$$

It follows that

$$\begin{aligned}
\frac{dC_\nu(d;\alpha)}{d\alpha} &= -\frac{(\sqrt{2\nu}d)^\nu}{\Gamma(\nu)2^{\nu-1}}\left[2\nu\alpha^{-\nu-1}K_\nu(\sqrt{2\nu}d\alpha^{-1}) - \alpha^{-\nu-2}(\sqrt{2\nu}d)K_{\nu+1}(\sqrt{2\nu}d\alpha^{-1})\right] \\
&= \frac{(\sqrt{2\nu}d)^{\nu+1}}{\Gamma(\nu)2^{\nu-1}\alpha^{\nu+2}}K_{\nu+1}(\sqrt{2\nu}d\alpha^{-1}) - \nu\frac{(\sqrt{2\nu}d)^\nu}{\Gamma(\nu)2^{\nu-2}\alpha^{\nu+1}}K_\nu(\sqrt{2\nu}d\alpha^{-1}) \\
&= \frac{(\sqrt{2\nu}d)^{\nu+1}}{\Gamma(\nu)2^{\nu-1}\alpha^{\nu+2}}K_{\nu+1}(\sqrt{2\nu}d\alpha^{-1}) - \frac{2\nu}{\alpha}C_\nu(d;\alpha) \\
&= \frac{2\nu}{\alpha}\frac{(\sqrt{2\nu}d)^{\nu+1}}{\Gamma(\nu+1)2^\nu\alpha^{\nu+1}}K_{\nu+1}(\sqrt{2\nu}d\alpha^{-1}) - \frac{2\nu}{\alpha}C_\nu(d;\alpha) \\
&= \frac{2\nu}{\alpha}\frac{\left(\sqrt{2(\nu+1)}\sqrt{\frac{\nu}{\nu+1}}d\right)^{\nu+1}}{\Gamma(\nu+1)2^\nu\alpha^{\nu+1}}K_{\nu+1}(\sqrt{2(\nu+1)}\sqrt{\frac{\nu}{\nu+1}}d\alpha^{-1}) - \frac{2\nu}{\alpha}C_\nu(d;\alpha) \\
&= \frac{2\nu}{\alpha}\left[C_{\nu+1}\left(\sqrt{\frac{\nu}{\nu+1}}d;\alpha\right) - C_\nu(d;\alpha)\right].
\end{aligned}$$

This completes the proof. ∎

This proposition provides a computational efficient approach to the first derivative of $C_v(d;\alpha)$ with respect to α in the score function.

2 Multivariate Probit Model

This section derives the composite estimating score function for multivariate probit model. Suppose we consider S spatial clusters and for $s = 1, \ldots, S$. Let $\boldsymbol{\iota}_s \in \mathbb{R}^d$ denote the spatial locations in a $d-$dimensional Euclidean space. We have n_s subjects nested in each cluster. Let y_{si} denote the binary outcome of the ith subject nested in cluster s, for $i = 1, \ldots, n_s$. We assume that

$$y_{si} = I[z_{si} > 0], \qquad \boldsymbol{Z} = (\mathbf{z}_1^\top, \ldots, \mathbf{z}_S^\top)^\top \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}), \tag{v}$$

where $\mathbf{z}_s = (z_{s1}, \ldots, z_{sn_s})^{\mathrm{T}}$ and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^{\mathrm{T}}, \ldots, \boldsymbol{\mu}_S^{\mathrm{T}})^{\mathrm{T}}$ with $\boldsymbol{\mu}_s = (\mu_{s1}, \ldots, \mu_{sn_s})^{\mathrm{T}}$. We assume that

$$\begin{aligned} \mu_{si} &= \Phi(\mathbf{x}_{si}^{\mathrm{T}}\boldsymbol{\beta}), \\ \boldsymbol{\Sigma}_{ss} &= \mathrm{var}[\mathbf{z}_s] = (1-\rho)\mathbf{I}_{n_s} + \rho \mathbf{1}_{n_s}\mathbf{1}_{n_s}^{\mathrm{T}}, \\ \boldsymbol{\Sigma}_{st} &= \mathrm{cov}[\mathbf{z}_s, \mathbf{z}_t^{\mathrm{T}}] = \rho C_\nu(d_{st}; \alpha)[\mathbf{1}_{n_s}\mathbf{1}_{n_t}^{\mathrm{T}}], \end{aligned} \quad \text{(vi)}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector and $\mathbf{x}_{si}$ is a $p \times 1$ vector, $d_{st} = \|\boldsymbol{\iota}_s - \boldsymbol{\iota}_t\|$ and $\|\cdot\|$ denotes an Euclidean distance. Functions $\Phi(x) = \int_{-\infty}^{x} \phi(t) \mathrm{d}t$ and $\phi(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)$ represent the standard normal distribution function and the standard normal density function respectively. Futhermore, if $(s,i) = (t,j)$, then $\sigma_{si,si} := \mathrm{var}(y_{si}) = \mu_{si} - \mu_{si}^2 = \Phi(\mathbf{x}_{si}^{\mathrm{T}}\boldsymbol{\beta})[1 - \Phi(\mathbf{x}_{si}^{\mathrm{T}}\boldsymbol{\beta})]$. If $(s,i) \neq (t,j)$, then

$$\begin{aligned} \sigma_{si,tj} &:= \mathrm{cov}(y_{si}, y_{tj}) = \pi_{si,tj} - \mu_{si}\mu_{tj} \\ &= \Phi_2(\mathbf{x}_{si}^{\mathrm{T}}\boldsymbol{\beta}, \mathbf{x}_{tj}^{\mathrm{T}}\boldsymbol{\beta}, \rho_{si,tj}) - \Phi(\mathbf{x}_{si}^{\mathrm{T}}\boldsymbol{\beta})\Phi(\mathbf{x}_{tj}^{\mathrm{T}}\boldsymbol{\beta}), \end{aligned} \quad \text{(vii)}$$

where $\pi_{si,tj} = \Phi_2(\mathbf{x}_{si}^{\mathrm{T}}\boldsymbol{\beta}, \mathbf{x}_{tj}^{\mathrm{T}}\boldsymbol{\beta}, \rho_{si,tj})$ and $\rho_{si,tj} = \rho C_\nu(d_{st}; \alpha)$.

The function $\Phi_2(x, y; \rho) = \int_{-\infty}^{x} \int_{-\infty}^{y} \phi_2(t, s; \rho) \mathrm{d}t\mathrm{d}s$ represents the standard bivariate normal distribution function, and the density function is

$$\phi_2(x, y; \rho) = \frac{1}{2\pi} \frac{1}{\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)}(x^2 + y^2 - 2\rho xy) \right\}.$$

denotes the standard bivariate normal density function.

Proposition 2.1 *The first order derivative of the standard bivariate normal distribution function with respect to the correlation parameter ρ is equal to the standard bivariate density function. i.e.*

$$\frac{d\Phi_2(x, y, \rho)}{d\rho} = \phi(x, y, \rho). \quad \text{(viii)}$$

Proof : First,

$$\frac{d\phi_2(x, y, \rho)}{d\rho} = \phi_2(x, y, \rho) \frac{\rho}{1-\rho^2} + \phi_2(x, y, \rho) \frac{(y - \rho x)(x - \rho y)}{(1-\rho^2)^2}.$$

Then,

$$\begin{aligned} \frac{d\Phi_2(x, y, \rho)}{d\rho} &= \Phi_2(x, y, \rho) \frac{\rho}{1-\rho^2} \\ &\quad + \frac{1}{(1-\rho^2)^2} \int_{-\infty}^{x} \int_{-\infty}^{y} \phi_2(u, v; \rho)(u - \rho v)(v - \rho u) dv du. \end{aligned} \quad \text{(ix)}$$

Also, by some routine linear algebra, we can show that

$$\int_{-\infty}^{x}\int_{-\infty}^{y} u^2\phi_2(u,v;\rho)dudv$$
$$= \left[\Phi_2(x,y;\rho)+\rho(1-\rho^2)\phi_2(x,y;\rho)\right] - \left[x\phi(x)\Phi\left(\frac{y-\rho x}{\sqrt{1-\rho^2}}\right)+\rho^2 y\phi(y)\Phi\left(\frac{x-\rho y}{\sqrt{1-\rho^2}}\right)\right],$$

$$\int_{-\infty}^{x}\int_{-\infty}^{y} v^2\phi_2(u,v;\rho)dudv$$
$$= \left[\Phi_2(x,y;\rho)+\rho(1-\rho^2)\phi_2(x,y;\rho)\right] - \left[y\phi(y)\Phi\left(\frac{x-\rho y}{\sqrt{1-\rho^2}}\right)+\rho^2 x\phi(x)\Phi\left(\frac{y-\rho x}{\sqrt{1-\rho^2}}\right)\right],$$

and

$$\begin{aligned} & \int_{-\infty}^{x}\int_{-\infty}^{y} uv\phi_2(u,v;\rho)dudv \\ = \; & (1-\rho^2)\phi_2(x,y;\rho) \\ & +\rho\left[\Phi_2(x,y;\rho) - x\phi(x)\Phi\left(\frac{y-\rho x}{\sqrt{1-\rho^2}}\right) - y\phi(y)\Phi\left(\frac{x-\rho y}{\sqrt{1-\rho^2}}\right)\right]. \end{aligned}$$

This further implies that

$$\begin{aligned} & \int_{-\infty}^{x}\int_{-\infty}^{y} \phi_2(u,v;\rho)(u-\rho v)(v-\rho u)dudv \\ & = (1-\rho^2)^2\phi_2(x,y;\rho) - \rho(1-\rho^2)\Phi_2(x,y;\rho). \end{aligned} \tag{x}$$

Combining (ix) and (x) completes the proof. ∎

Also by some routine algebra, we obtain

$$\frac{d\Phi_2(x,y;\rho)}{dx} = \Phi\left(\frac{y-\rho x}{\sqrt{1-\rho^2}}\right)\phi(x), \qquad \frac{d\Phi_2(x,y;\rho)}{dy} = \Phi\left(\frac{x-\rho y}{\sqrt{1-\rho^2}}\right)\phi(y). \tag{xi}$$

From equations (viii) and (xi), the general score function is given by

$$\mathbf{U}_{si,tj}(\boldsymbol{\theta}) \;\; = \;\; \boldsymbol{\Delta}_{si,tj}^{T}\mathbf{V}_{si,tj}^{-1}\mathbf{R}_{si,tj},$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \alpha, \rho)^{\mathrm{T}}$. For $(s,i) \neq (t,j)$,

$$\boldsymbol{\Delta}_{si,tj}(\boldsymbol{\theta}) = \frac{d}{d\boldsymbol{\theta}}\begin{pmatrix} \mu_{si} \\ \mu_{tj} \\ \sigma_{si,tj} \end{pmatrix} = \begin{pmatrix} \frac{d\mu_{si}}{d\boldsymbol{\beta}^{\mathrm{T}}} & \frac{d\mu_{si}}{d\alpha} & \frac{d\mu_{si}}{d\rho} \\ \frac{d\mu_{tj}}{d\boldsymbol{\beta}^{\mathrm{T}}} & \frac{d\mu_{tj}}{d\alpha} & \frac{d\mu_{tj}}{d\rho} \\ \frac{d\sigma_{si,tj}}{d\boldsymbol{\beta}^{\mathrm{T}}} & \frac{d\sigma_{si,tj}}{d\alpha} & \frac{d\sigma_{si,tj}}{d\rho} \end{pmatrix}$$

Specifically,

$$
\begin{aligned}
\frac{d\mu_{si}}{d\boldsymbol{\beta}^{\mathrm{T}}} &= \phi(\mathbf{x}_{si}^{\mathrm{T}}\boldsymbol{\beta})\mathbf{x}_{si}^{\mathrm{T}} \\
\frac{d\mu_{tj}}{d\boldsymbol{\beta}^{\mathrm{T}}} &= \phi(\mathbf{x}_{tj}^{\mathrm{T}}\boldsymbol{\beta})\mathbf{x}_{tj}^{\mathrm{T}} \\
\frac{d\sigma_{si,tj}}{d\boldsymbol{\beta}^{\mathrm{T}}} &= \frac{d\Phi_2(\mathbf{x}_{si}^{\mathrm{T}}\boldsymbol{\beta}, \mathbf{x}_{tj}^{\mathrm{T}}\boldsymbol{\beta}; \rho_{si,tj})}{d\boldsymbol{\beta}^{\mathrm{T}}} - \frac{d\Phi(\mathbf{x}_{si}^{\mathrm{T}}\boldsymbol{\beta})\Phi(\mathbf{x}_{tj}^{\mathrm{T}}\boldsymbol{\beta})}{d\boldsymbol{\beta}^{\mathrm{T}}} \\
\frac{d\mu_{si}}{d\alpha} &= 0 \\
\frac{d\mu_{tj}}{d\alpha} &= 0 \\
\frac{d\sigma_{si,tj}}{d\alpha} &= \frac{2\rho}{\alpha}\left[(\nu+1)C_{\nu+1}\left(\sqrt{\frac{\nu}{\nu+1}}d_{st}; \alpha\right) - \nu C_\nu(d_{st}; \alpha)\right]\phi_2(\mathbf{x}_{si}^{\mathrm{T}}\boldsymbol{\beta}, \mathbf{x}_{tj}^{\mathrm{T}}\boldsymbol{\beta}, \rho_{si,tj}) \\
\frac{d\mu_{si}}{d\rho} &= 0 \\
\frac{d\mu_{tj}}{d\rho} &= 0 \\
\frac{d\sigma_{si,tj}}{d\rho} &= C_\nu(d_{st}; \alpha)\phi_2(\mathbf{x}_{si}^{\mathrm{T}}\boldsymbol{\beta}, \mathbf{x}_{tj}^{\mathrm{T}}\boldsymbol{\beta}, \rho_{si,tj})
\end{aligned}
$$

where

$$
\begin{aligned}
\frac{d\Phi_2(\mathbf{x}_{si}^{\mathrm{T}}\boldsymbol{\beta}, \mathbf{x}_{tj}^{\mathrm{T}}\boldsymbol{\beta}, \rho)}{d\boldsymbol{\beta}^{\mathrm{T}}} &= \mathbf{x}_{si}^{\mathrm{T}}\nabla_1\Phi_2(\mathbf{x}_{si}^{\mathrm{T}}\boldsymbol{\beta}, \mathbf{x}_{tj}^{\mathrm{T}}\boldsymbol{\beta}, \rho_{si,tj}) + \mathbf{x}_{tj}^{\mathrm{T}}\nabla_2\Phi_2(\mathbf{x}_{si}^{\mathrm{T}}\boldsymbol{\beta}, \mathbf{x}_{tj}^{\mathrm{T}}\boldsymbol{\beta}, \rho_{si,tj}) \\
&= \mathbf{x}_{si}^{\mathrm{T}}\phi(\mathbf{x}_{si}^{\mathrm{T}}\boldsymbol{\beta})\Phi\left(\frac{\mathbf{x}_{tj}^{\mathrm{T}}\boldsymbol{\beta} - \rho_{si,tj}\mathbf{x}_{si}^{\mathrm{T}}\boldsymbol{\beta}}{\sqrt{1-\rho_{si,tj}^2}}\right) + \mathbf{x}_{tj}^{\mathrm{T}}\phi(\mathbf{x}_{tj}^{\mathrm{T}}\boldsymbol{\beta})\Phi\left(\frac{\mathbf{x}_{si}^{\mathrm{T}}\boldsymbol{\beta} - \rho_{si,tj}\mathbf{x}_{tj}^{\mathrm{T}}\boldsymbol{\beta}}{\sqrt{1-\rho_{si,tj}^2}}\right),
\end{aligned}
$$

where $\nabla_r\Phi_2(x_1, x_2, \rho)$ is the first order derivative of Φ_2 with respect to argument $x_r, r = 1, 2$. Also,

$$
\frac{d\Phi(\mathbf{x}_{si}^{\mathrm{T}}\boldsymbol{\beta})\Phi(\mathbf{x}_{tj}^{\mathrm{T}}\boldsymbol{\beta})}{d\boldsymbol{\beta}^{\mathrm{T}}} = \mathbf{x}_{si}^{\mathrm{T}}\phi(\mathbf{x}_{si}^{\mathrm{T}}\boldsymbol{\beta})\Phi(\mathbf{x}_{tj}^{\mathrm{T}}\boldsymbol{\beta}) + \mathbf{x}_{tj}^{\mathrm{T}}\phi(\mathbf{x}_{tj}^{\mathrm{T}}\boldsymbol{\beta})\Phi(\mathbf{x}_{si}^{\mathrm{T}}\boldsymbol{\beta}).
$$

Note that

$$
\mathbf{R}_{si,tj} = \begin{pmatrix} y_{si} - \mu_{si} \\ y_{tj} - \mu_{tj} \\ (y_{si} - \mu_{si})(y_{tj} - \mu_{tj}) - \sigma_{si,tj} \end{pmatrix}.
$$

And

$$\mathbf{V}_{si,tj} = \mathrm{Var}(\mathbf{R}_{si,tj}) = \begin{pmatrix} \mu_{si}(1-\mu_{si}) & \sigma_{si,tj} & \sigma_{si,tj}(1-2\mu_{si}) \\ \sigma_{si,tj} & \mu_{tj}(1-\mu_{tj}) & \sigma_{si,tj}(1-2\mu_{tj}) \\ \sigma_{si,tj}(1-2\mu_{si}) & \sigma_{si,tj}(1-2\mu_{tj}) & [\boldsymbol{V}_{si,tj}]_{3,3} \end{pmatrix},$$

where

$$[\boldsymbol{V}_{si,tj}]_{3,3} = \pi_{si,tj}(1-\pi_{si,tj}) + \pi_{si,tj}[6\mu_{si}\mu_{tj} - 2\mu_{si} - 2\mu_{tj}] + 2\mu_{si}\mu_{tj}(\mu_{si} + \mu_{tj} - 4\mu_{si}\mu_{tj}).$$

3 Large Sample Properties

3.1 Definition of the Distance Metric ϱ

Consider the paired random process of the following form:

$$\boldsymbol{y}(k) \equiv \left(y_{si}, y_{tj}\right)^{\mathrm{T}}, \;\; k = (s,i,t,j) \in \mathcal{D}_n. \tag{xii}$$

The distance between $\boldsymbol{y}(k_1)$ and $\boldsymbol{y}(k_2)$, $k_1, k_2 \in \mathcal{D}_n$, depends on configurations of four points in the spatial-clustered domain $\mathbb{R}^2 \times \mathbb{Z}$. The coordinates of one point is denoted by (s,i), where s is the vector of spatial coordinates, and i is the index of subject within a cluster. The distance between two points $p_1 = (s,i)$ and $p_2 = (t,j)$ in $\mathbb{R}^2 \times \mathbb{Z}$ is defined as $\xi(p_1, p_2) = ||s-t|| + I(i \neq j)d_0$, where $|| \bullet ||$ is the Euclidean distance in $\mathcal{R}^2$. Defined in this way, the distance between any two different observations consists of two parts. The first part is the spatial distance between two clusters they reside in, and the second part is d_0 if they have different indices within clusters. This ensures that different observations are at least d_0 distance away. Moreover the distance between two points $k_1 = (p_1, p_2)$ and $k_2 = (p_1', p_2')$ in $\mathcal{D}_n \subset \mathbb{R}^2 \times \mathbb{Z} \times \mathbb{R}^2 \times \mathbb{Z}$ is defined as

$$\varrho(k_1, k_2) = \min\{\xi(p_1, p_2), \xi(p_1, p_2'), \xi(p_1', p_2), \xi(p_1', p_2')\}, \tag{xiii}$$

that is, the minimum distance of two points in sets (p_1, p_2) and (p_1', p_2'). The distance between any subsets $D_1, D_2 \subset \mathcal{D}_n$ is defined as

$$\varrho(D_1, D_2) = \min\{\varrho(k_1, k_2) : k_1 \in D_1, k_2 \in D_2\}. \tag{xiv}$$

Mixing Conditions for the paired process $\{\boldsymbol{y}(k), k \in \mathcal{D}_n\}$

To regulate the dependence structure of $\boldsymbol{y}(k)$ defined in equation (xii), we impose some α-mixing conditions on $\boldsymbol{y}(k)$. Let D_1 and D_2 be two subsets of $\mathcal{D}_n$, and let $\sigma(D_1) = \sigma\{\boldsymbol{y}(k); k \in D_1\}$ be the σ-algebra generated by random variables $\{\boldsymbol{y}(k), k \in D_1\}$. Define

$$\alpha(D_1, D_2) = \sup\{|P(A \cap B) - P(A)P(B)|; A \in \sigma(D_1), B \in \sigma(D_2)\}.$$

Then this α-mixing coefficient for the random field $\{\boldsymbol{y}(k), k \in D_n\}$ is defined as:

$$\alpha(k, l, m) = \sup\{\alpha(D_1, D_2), |D_1| < k, |D_2| < l, \varrho(D_1, D_2) \geq m\},$$

with $k, l, m \in \mathbb{N}$ and $\varrho(D_1, D_2)$ the distance between sets D_1 and D_2 defined in (xiv). We need the following conditions similar to those stated in Assumption 3 by Jenish and Prucha (2009).

Mixing Conditions The process $\{\boldsymbol{y}(k), k \in D_n\}$ satisfies the following mixing conditions in an a-dimensional space:

(a) $\sum_{m=1}^{\infty} m^{a-1}\alpha(1, 1, m)^{\delta/(2+\delta)} < \infty$, for some $\delta > 0$,

(b) $\sum_{m=1}^{\infty} m^{a-1}\alpha(k, l, m) < \infty$ for $k + l \leq 4$,

(c) $\alpha(1, \infty, m) = O(m^{-a-\epsilon})$ for some $\epsilon > 0$.

It suffices to require a polynomial decay of the α-mixing coefficient, which can be shown to hold for the Matérn class.

3.2 Detailed Derivation

In the spatial-clustered setting considered here, the increase of the sample size can be achieved by either increasing the number of subjects within each cluster, or by increasing the number of spatial clusters. For the latter case, two scenarios are possible: (i) more sample locations are added within a fixed spatial domain, known as the in-fill asymptotics (Zhang, 2004); and (ii) more locations are included by expanding the spatial domain, corresponding to the increasing-domain asymptotics (Mardia and Marshall, 1984). Sampling more people within clusters can be regarded as a special case under the in-fill asymptotic scenario, where more observations are collected at the same locations. Since these extra data are likely to be highly correlated, for some parameters, consistent estimates may not exist under the in-fill asymptotics (Zhang, 2004). In this paper, we establish large-sample properties under the increasing domain context.

Under appropriate conditions of correlation decay rates for the process $\boldsymbol{y}(k)$, usually postulated by certain mixing conditions (Guyon, 1995), we expect to have "pseudo-independent" pairs when they are beyond a certain distance. In such cases, we can derive laws of large numbers (LLN) and central limit theorems for $\Psi_{B,n}(\boldsymbol{\theta}, d)$ and $\Psi_{W,n}(\boldsymbol{\theta})$ respectively, and then for $\boldsymbol{\Gamma}_n(\boldsymbol{\theta}, d) = \left(\Psi^{\mathrm{T}}_{B,n}(\boldsymbol{\theta}, d), \Psi^{\mathrm{T}}_{W,n}(\boldsymbol{\theta})\right)^{\mathrm{T}}$. For notation simplicity, we omit d in the expression, i.e. denoted $\boldsymbol{\Gamma}_n(\boldsymbol{\theta}) = \boldsymbol{\Gamma}_n(\boldsymbol{\theta}, d)$ and $\Psi_{B,n}(\boldsymbol{\theta}) = \Psi_{B,n}(\boldsymbol{\theta}, d)$, and so on. Furthermore, by using the standard GMM arguments (Hansen, 1982), we can show the consistency and asymptotic normality of the JCEF estimator $\widehat{\boldsymbol{\theta}}_n$ defined in equation (5) in the paper.

Jenish and Prucha (2009) developed a set of limit theorems for random processes under rather general conditions of nonstationarity, unevenly spaced locations, and general forms of sample regions. We exploit those results to sketch large-sample properties for our JCEF estimator as follows.

Consistency Consider a generic case of composite estimating function

$$\Psi_{\mathcal{A},n}(\boldsymbol{\theta}) = \frac{1}{|\mathcal{D}_{\mathcal{A},n}|} \sum_{k \in \mathcal{D}_{\mathcal{A},n}} \boldsymbol{U}_k(\boldsymbol{y}(k); \boldsymbol{\theta}), \;\; \mathcal{A} \in \{B, W\}.$$

We assume the following assumptions for the component score functions.

Assumption 1 The (possibly unevenly spaced) lattice $D \subset R^2 \times Z^+ \times R^2 \times Z^+$ is infinitely countable. All elements in D are located at distances of at least $d_0 > 0$ from each other. That is, $\varrho(k_1, k_2) \geq d_0$, for all $k_1, k_2 \in D$, where $\varrho(k_1, k_2)$ is a distance metric for any two points $k_1, k_2 \in D$ defined in (xiii).

Assumption 2 $\{D_{\mathcal{A},n} : n \in N\}$ is a sequence of arbitrary finite subsets of D, satisfying $|D_{\mathcal{A},n}| \to \infty$ as $n \to \infty$, for $\mathcal{A} \in \{B, W\}$.

Assumption 3 (Uniform $L_{2+\delta}$ integrability) Let $Q_k = \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} ||\boldsymbol{U}_k(\boldsymbol{y}(k); \boldsymbol{\theta})||$. For some $\delta > 0$, $\lim_{e\to\infty} EQ_k^{2+\delta}\mathbf{1}(|Q_k| > e) = 0$, for all $k \in D_n$, where $D_n \equiv D_{B,n} \cup D_{W,n}$.

Assumption 4 $E \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} ||\nabla_{\boldsymbol{\theta}} \boldsymbol{U}_k(\boldsymbol{y}(k); \boldsymbol{\theta})|| < \infty$, for all $k \in D_n$.

Assumption 1 ensures that the increase of sample size is achieved by an expanding domain, thus it rules out the in-fill asymptotics. Assumption 2 guarantees that sequences of subsets $D_{B,n}$ and $D_{W,n}$ on which the process is generated, increases in cardinality. Assumptions 3 and 4 are regularity conditions for the composite score functions. The uniform integrability condition in Assumption 3 is a standard moment assumption required by the CLTs for one-dimensional processes. A sufficient condition for the uniform $L_{2+\delta}$ integrability of $\boldsymbol{U}_k$ is its uniform L_γ boundedness for some $\gamma > 2 + \delta$. A weaker assumption of L_1 integrability is sufficient for an LLN on $\boldsymbol{U}_k$. Assumption 4 is a Lipschitz-type condition, implying that the composite score functions are L_0 stochastically equicontinuous, so that a uniform law of the large numbers (ULLN) can be established.

Lemma 3.1 *When the sample size increases with the increasing spatial domain, under assumptions 1 - 4 above, and the appropriate mixing conditions (a)-(b) for the process $\{\boldsymbol{y}(k), k \in \mathcal{D}_n\}$ given in Section 3.1, we have* $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} ||\Psi_{\mathcal{A},n}(\boldsymbol{\theta}) - E\Psi_{\mathcal{A},n}(\boldsymbol{\theta})|| \xrightarrow{p} 0,$ *as* $n \to \infty$.

As shown in Jenish and Prucha (2009), a polynomial decay of the mixing coefficient for the process is required, which is satisfied by the Matérn spatial covariance considered in this paper, since the decay order of the Matérn spatial correlation is governed by the term of $\exp(-d/\alpha)$ (Dempsey and Benson, 1960).

Lemma 3.1 holds for $\Psi_{B,n}(\boldsymbol{\theta}, d)$ and $\Psi_{W,n}(\boldsymbol{\theta})$ respectively, so we can show easily that for any given positive-definite weight matrix W in equation (5) in the paper,

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |Q_n(\boldsymbol{\theta}) - EQ_n(\boldsymbol{\theta})| \xrightarrow{p} 0, \quad \text{as } n \to \infty.$$

Consequently, we establish the consistency of the JCEF estimator in Theorem 1.

Theorem 3.2 *Under the same regularity conditions stated in Lemma 3.1, if the true parameter value* $\boldsymbol{\theta}_0$ *is the unique minimizer of* $EQ_n(\boldsymbol{\theta})$*, and* $\hat{\boldsymbol{\theta}}_n$ *minimizes* $Q_n(\boldsymbol{\theta})$*, then* $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0, \quad \text{as } n \to \infty.$

Asymptotic Normality To derive the asymptotic distribution of the JCEF estimator, the following additional regularity conditions are needed.

Assumption 5 Let $\boldsymbol{W}_n^*(\boldsymbol{\theta}) = \text{Var}\{\boldsymbol{\Gamma}_n(\boldsymbol{\theta})\}$. Assume $\lim_{n\to\infty} n\boldsymbol{W}_n^*(\boldsymbol{\theta}) = \Lambda(\boldsymbol{\theta})$, where $\Lambda(\boldsymbol{\theta})$ is a positive-definite matrix.

Assumption 6 $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} ||\nabla_{\boldsymbol{\theta}}\boldsymbol{\Gamma}_n(\boldsymbol{\theta}) - E\nabla_{\boldsymbol{\theta}}\boldsymbol{\Gamma}_n(\boldsymbol{\theta})|| \xrightarrow{p} 0$. Write $\lim_{n\to\infty} E\nabla_{\boldsymbol{\theta}}\boldsymbol{\Gamma}_n(\boldsymbol{\theta}) = \boldsymbol{I}(\boldsymbol{\theta})$, where $\boldsymbol{I}(\boldsymbol{\theta})$ is a positive-definite information matrix.

Assumption 5 assumes the variance of $\boldsymbol{\Gamma}_n(\boldsymbol{\theta})$ is of order $O(n^{-1})$, a standard assumption for covariance subsampling estimation. Assumption 6 is the condition for ULLN required by the Hessian matrix $\nabla_{\boldsymbol{\theta}}\boldsymbol{\Gamma}_n(\boldsymbol{\theta})$, which regulates the asymptotic variance of the estimator and can be obtained with the same regularity conditions in Lemma 3.1.

Lemma 3.3 *Under increasing domain framework, given Assumptions 1-6 above, we have* $\sqrt{n}\,\boldsymbol{\Gamma}_n(\boldsymbol{\theta}) \xrightarrow{L} N(\mathbf{0}, \Lambda(\boldsymbol{\theta})), \text{ as } n \to \infty.$

A sketch of the proof for Lemma 3.3 is given in the Appendix of Bai et al. (2012). Then using the standard GMM arguments (Hansen, 1982), we establish the following theorem:

Theorem 3.4 *Under the increasing domain framework, given Assumptions 1-6 above and the mixing conditions (a)-(b) in Section 3.1 for the process* $\{\boldsymbol{y}(k), k \in \mathcal{D}_n\}$*, we have*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{L} N(0, \Omega(\boldsymbol{\theta}_0)\Lambda(\boldsymbol{\theta}_0)\Omega^{\mathrm{T}}(\boldsymbol{\theta}_0)), \text{ as } n \to \infty,$$

where $\Omega(\boldsymbol{\theta}_0) = -\{\boldsymbol{I}^{\mathrm{T}}(\boldsymbol{\theta}_0)\Lambda^{-1}(\boldsymbol{\theta}_0)\boldsymbol{I}(\boldsymbol{\theta}_0)\}^{-1}\boldsymbol{I}^{\mathrm{T}}(\boldsymbol{\theta}_0)\Lambda^{-1}(\boldsymbol{\theta}_0).$

4 R Package GeoCopula

A user-friendly `R` package `GeoCopula` is available under the Paper Information link at the *Biometrics* webpage `http://www.biometrics.tibs.org`. This R package supplies both MAC OS X and DOS Window OS versions, and can be directly downloaded from

`http://web1.sph.emory.edu/users/jkang30/software/GeoCopula.html`

It provides an efficient composite likelihood estimation and inference in the analysis of spatial-clustered continuous and binary data using GeoCopula regression models, where point estimation and inference are implemented in two separate functions for the sake of computing time. In the point estimation function the weight matrix $\boldsymbol{W}^*$ is specified by either a sample covariance matrix or an estimate from the subsampling method. In the inference function both parametric bootstrap and subsampling methods are available to obtain standard error estimate and 95% confident interval.

To run functions in the `GeoCopula` package, one must first upload `R` package `mvtnorm`. Below is an example of running the `GeoCopula` package:

```
##load R package mvtnorm before load GeoCopula##

my.seed = 2015

model = "binary"
para = list(beta=c(1,1,-1),alpha=2.0,rho=0.8,nu=1.5)
numOfClusters= 100
numOfSubjects = rep(4,length=numOfClusters)
spatialGrids=expand.grid(1:20,1:20)
sigmaX=3

### Simulate Data
dat = geoCopula.simulate.data(numOfClusters=numOfClusters, numOfSubjects = numOfSubjects,
        spatialGrids = spatialGrids,model=model,para=para,seed=my.seed,sigmaX=sigmaX)
```

```
trucdist=4.0
between.indices = geoCopula.compute.between.cluster.indices(dat$dist,dat$clusterPairIndices,
              dat$numOfClusters,dat$clusterSubjIndices,trucdist)

within.indices =
geoCopula.compute.within.cluster.indices(dat$numOfClusters,dat$clusterSubjIndices)

### Sample Subregions

sub.region.size = 4
numOfSubRegions = 20
subregion.indices=geoCopula.sample.subregion(dat$locations,sub.region.size=sub.region.size,
                              numOfSubRegions=numOfSubRegions,my.seed=my.seed)
subregion=geoCopula.subregion.index.pair(subregion.indices,dat$subjClusterIndices,
                  within.indices,between.indices)

### Compute initial parameter estimates
nu = 1.5
para0 = geoCopula.initial.parameter(dat=dat,model=model,nu=nu,
             between.indices=between.indices,within.indices=within.indices)

### Estimate parameter using GeoCopula given pre-specified nu
approach = "simple"
weighted = TRUE
method="Nelder-Mead"
reltol=1e-1
trace=TRUE
maxit=5000

para.est = geoCopula.parameter.estimate(dat=dat,model=model,para0=para0,
                 between.indices=between.indices,within.indices=within.indices,
```

```
                 method=method,reltol=reltol,trace=trace,
                 maxit=maxit,weighted=weighted,approach=approach,subregion=subregion)

### Profile QIF estimate
nulist = seq(0.5,3.0,by=0.5)

res = geoCopula.parameter.estimate.nu(dat=dat,model=model,nulist=nulist,
         between.indices=between.indices,within.indices=within.indices,
         method=method,reltol=reltol,trace=trace,maxit=maxit,
         weighted=weighted,approach=approach,subregion=subregion)

plot(nulist,res$nuQfun,type="b",xlab=expression(nu),ylab="Q")
print(res$para.est)

## Make inference on the model parameters
se.approach="sub.sample"
numOfBootstrap=200
se.numOfSubRegions=100
se.sub.region.size=4

inference=geoCopula.inference(dat=dat, model=model, para.est=para.est,
                 se.approach=se.approach, se.numOfSubRegions=se.numOfSubRegions,
                 se.sub.region.size=se.sub.region.size, numOfBootstrap=numOfBootstrap,
                 between.indices=between.indices,within.indices=within.indices,
                 method=method,reltol =reltol, trace=trace,maxit=maxit,
                 weighted=weighted,my.seed=my.seed,
                 approach="simple",subregion=NULL)

print(geoCopula.summary.inference(inference))
```

References

Bai, Y., Song, P. X.-K., and Raghunathan, T. (2012). Joint composite estimating functions in spatiotemporal models. *Journal of the Royal Statistical Society, Series B* **74,** 799–824.

Dempsey, E. and Benson, G. (1960). Note on the asymptotic expansion of the modified bessel function of the second kind. *Mathematics of Computation* **14,** 362–365.

Guyon, X. (1995). *Random fields on a network: modeling, statistics and applications*. Springer-Verlag.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50,** 1029–1054.

Jenish, N. and Prucha, I. R. (2009). Central limit theorems and uniform laws of large numbers for arrays of random fields. *Journal of Econometrics* **150,** 86–98.

Mardia, K. and Marshall, R. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71,** 135–146.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association* **99,** 250–261.