GLS-SOD: A Generalized Local Statistical Approach for Spatial Outlier Detection

Feng Chen, Chang-Tien Lu, Arnold P. Boedihardjo
Virginia Tech, Computer Science Department

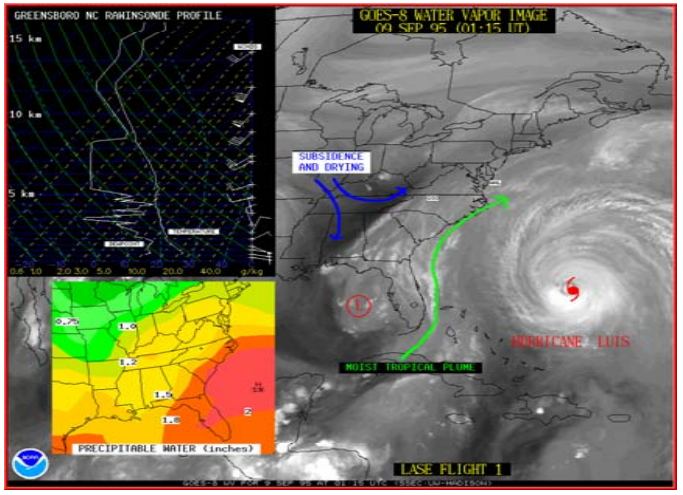
July, 28, 2010

KDD ’2010, Washington, DC, USA

Outline

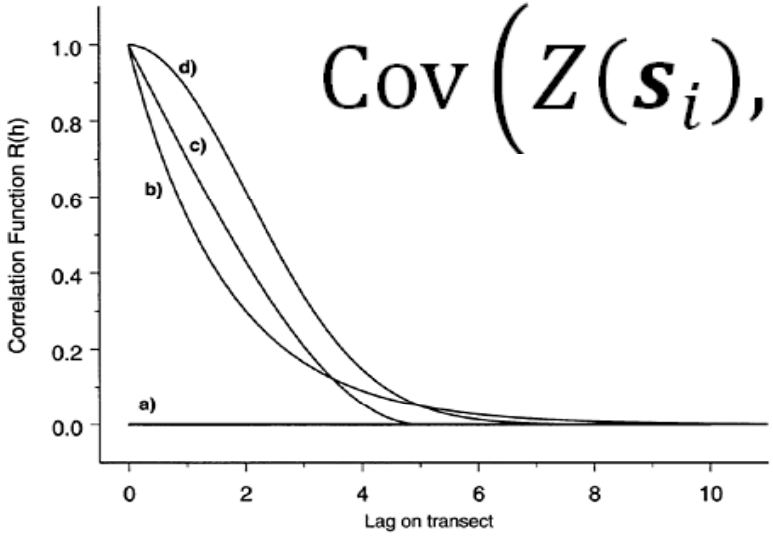
- Motivations
- Generalized Local Statistical (GLS) Model
- GLS Robust Estimation & Inferences
- Simulations
- Summary

Applications



What's Special for Spatial Data?

- By the first law of Geography, "*Everything is related to everything else, but nearby things are more related than distant things*" [Tobler 79]
- Spatial autocorrelation
  - Correlation of a variable with itself through space



$$\mathrm{Cov}\left(Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)\right) = f\left(\left\|\boldsymbol{s}_i - \boldsymbol{s}_j\right\|; \boldsymbol{\theta}\right)$$

Global based Spatial Outlier Detection

Given a partial sample $\{Z(\boldsymbol{s}_i)\}_{i=1}^{n}$ of a Gaussian random field, let

$\boldsymbol{Z} = [Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n)]^T, \boldsymbol{X} = [\boldsymbol{x}(\boldsymbol{s}_1), \ldots, \boldsymbol{x}(\boldsymbol{s}_n)]^T$

$$\boldsymbol{Z} \sim N(\boldsymbol{X\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$$

$$\min_{\boldsymbol{\beta},\boldsymbol{\theta}} \{(\boldsymbol{Z} - \boldsymbol{X\beta})^T [\boldsymbol{\Sigma}(\boldsymbol{\theta})]^{-1} (\boldsymbol{Z} - \boldsymbol{X\beta})\}$$

subject to some constraints on θ.

Standardized Z Test for an observation $Z(\boldsymbol{s})$.

First calculate $E[Z(\boldsymbol{s})] = \boldsymbol{x}(\boldsymbol{s})^T\boldsymbol{\beta}$, $\mathrm{Var}(Z(\boldsymbol{s})) = \sigma_0^2$, $\mathrm{Cov}[\boldsymbol{Z}, Z(\boldsymbol{s})] = \boldsymbol{\sigma}$, $\mathrm{Var}[\boldsymbol{Z}] = \boldsymbol{\Sigma}$, and $E[\boldsymbol{Z}(\boldsymbol{s})] = \boldsymbol{\mu}$. Then

$$\frac{Z(\boldsymbol{s}) - \boldsymbol{x}(\boldsymbol{s})^T\boldsymbol{\beta} - \boldsymbol{\sigma}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Z} - \boldsymbol{X\beta})}{\sqrt{\sigma_0^2 - \boldsymbol{\sigma}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}^T}} \geq \Phi^{-1}\left(\frac{\alpha}{2}\right), \quad e.x.\ \alpha = 0.05$$

Local based Spatial Outlier Detection

$$Z \sim N(X\beta, \Sigma(\theta))$$

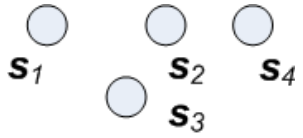
Local (laplacian) smoothing

$$FZ \sim N(FX\beta, F\Sigma(\theta)F^T)$$

$$FZ \sim N(\mu \cdot I_{n\times 1}, \sigma^2 I_{n\times n})$$

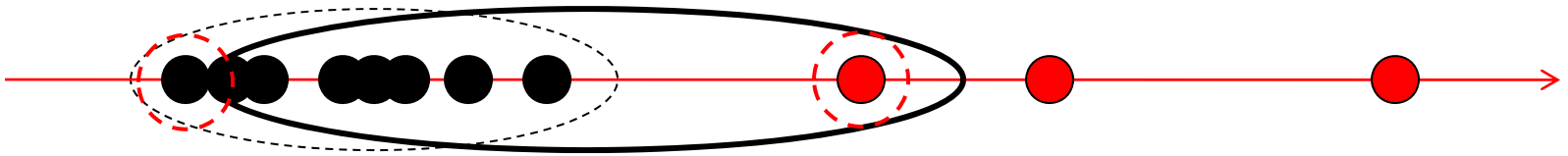
$$FX\beta \approx \mu \cdot I_{n\times 1}$$

$$F\Sigma(\theta)F^T \approx \sigma^2 I_{n\times n}$$



$$F = \begin{bmatrix} 1 & -½ & -½ & 0 \\ 0 & 1 & -½ & -½ \\ -½ & -½ & 1 & 0 \\ 0 & -½ & -½ & 1 \end{bmatrix}$$

Z Test: $\frac{(Z(s)-\mu)}{\sigma} \geq \Phi^{-1}\left(\frac{\alpha}{2}\right)$, $e.x.\ \alpha = 0.05$



swamping masking

Why GLS-SOD?

- Global based
  - Pros: high accuracy with statistical justifications.
  - Cons: very slow; complicate estimation process; non-convex optimization
- Local based
  - Pros: very fast; simplicity
  - Cons: heuristic-driven, lack of statistical justifications
- Questions
  - Local (laplacian) smoothing vs. spatial dependence?
  - Statistical connections between local and global methods?
  - When will existing local based methods perform poorly and how to handle these situations?

Gaussian Random Field

Consider a Gaussian Random Field $\{Z(\boldsymbol{s}), \boldsymbol{s} \in D \subset \mathbb{R}^2\}$

$$Z(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})^T\boldsymbol{\beta} + \omega(\boldsymbol{s}) + \epsilon(\boldsymbol{s})$$

Given a partial sample $\{Z(\boldsymbol{s}_i)\}_{i=1}^n$ of the random field, let

$$\boldsymbol{Z} = [Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n)]^T, \boldsymbol{X} = [\boldsymbol{x}(\boldsymbol{s}_1), \ldots, \boldsymbol{x}(\boldsymbol{s}_n)]^T$$

$$\boldsymbol{e} = [e(\boldsymbol{s}_1), \ldots e(\boldsymbol{s}_n)]^T, \boldsymbol{\omega} = [\omega(\boldsymbol{s}_1), \ldots \omega(\boldsymbol{s}_n)]^T$$

Large scale trend

White noise variation

$$\boldsymbol{Z} = \boldsymbol{X\beta} + \boldsymbol{\omega} + \boldsymbol{e} \sim N(\boldsymbol{X\beta}, \boldsymbol{\Sigma} + \sigma_0^2\boldsymbol{I})$$

Small Scale Variation

Generalized Local Statistical Model

$$Z = X\beta + \omega + e \sim N(X\beta, \Sigma + \sigma_0^2 I)$$

Generalized local statistical model (GLS)

$$\mathrm{diff}(Z) = FZ \sim N(FX\beta, F\Sigma F^T + \sigma_0^2 FF^T)$$

$\approx$ (See theorem 3)

$$\mathrm{diff}(Z) = FZ \sim N(FX\beta, \sigma^2 I + \sigma_0^2 FF^T)$$

Convolution effects

When neighbor size $\geq 10, \sigma_0^2 FF^T \approx \sigma_0^2 I$ (See Theorem 1)

Generalized least Squares

$$\mathbf{diff}(\boldsymbol{Z}) \sim \mathbf{N}(\boldsymbol{FX\beta}, \sigma^2\boldsymbol{I} + \sigma_0^2\boldsymbol{FF}^T)$$

$$\arg\min_{\beta,\sigma_0,\sigma} \left[(\boldsymbol{FZ} - \boldsymbol{FX\beta})^T\left(\sigma^2\boldsymbol{I} + \sigma_0^2\boldsymbol{FF}^T\right)^{-1}(\boldsymbol{FZ} - \boldsymbol{FX\beta})\right],$$

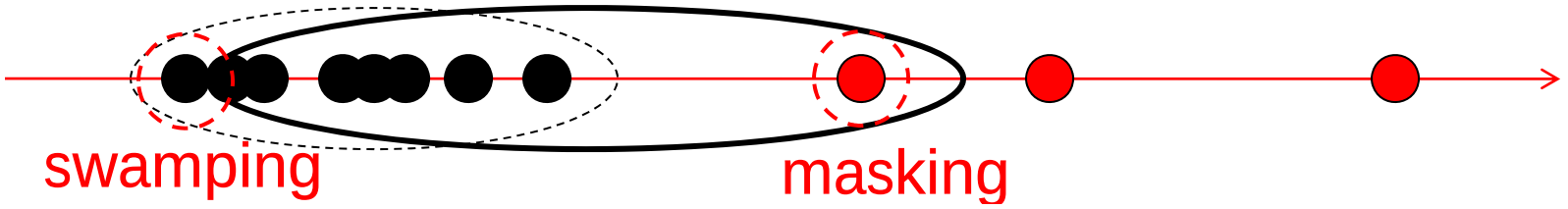
$$\text{subject to } \sigma_0^2 + \sigma^2 = 1 \text{ and } \sigma_0, \sigma \geq 0.$$

$$\arg\min_{\boldsymbol{\beta},\sigma_0,\sigma} \left[\sum_{i=1}^{n} \frac{\{(\boldsymbol{FZ} - \boldsymbol{FX\beta})^T\boldsymbol{q}_i\}^2}{\sigma^2 + \sigma_0^2\lambda_i}\right], \text{s. t. } \sigma_0^2 + \sigma^2 = 1;\ \sigma_0, \sigma \geq 0.$$

$$\frac{\partial^2 f_i}{\partial \boldsymbol{\theta}^2} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{F}\boldsymbol{q}_i(\sigma^2 + \sigma_0^2\lambda_i) \\ \left(\boldsymbol{q}_i^T Z - \boldsymbol{q}_i^T\boldsymbol{FX\beta}\right)^T \\ \lambda_i\left(\boldsymbol{q}_i^T Z - \boldsymbol{q}_i^T\boldsymbol{FX\beta}\right)^T \end{bmatrix} \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{F}\boldsymbol{q}_i(\sigma^2 + \sigma_0^2\lambda_i) \\ \left(\boldsymbol{q}_i^T\boldsymbol{Z} - \boldsymbol{q}_i^T\boldsymbol{FX\beta}\right)^T \\ \lambda_i\left(\boldsymbol{q}_i^T\boldsymbol{Z} - \boldsymbol{q}_i^T\boldsymbol{FX\beta}\right)^T \end{bmatrix}^T \succcurlyeq 0.$$

Forward / Backward Search

- GLS Backward Search
  - Model estimation by generalized least squares
  - Remove the most probable outlier and update all local differences
  - Repeat until the p values of all existing objects are greater than a threshold (e.g., 0.025)
- GLS Forward Search
  - Estimation by a robust subset **S** of local differences
  - Add test objects one by one to the training set **S**
  - Check the change of the smallest p value in **S**
  - A large drop in the smallest p value indicates an outlier



GLS Z-Test Statistics

Test if an observation $Z(\boldsymbol{s}_n)$ is abnormal:

$$\left\{[\sigma^2\boldsymbol{I} + \sigma_0^2\boldsymbol{F}\boldsymbol{F}^{\boldsymbol{T}}]^{-\frac{\mathbf{1}}{\mathbf{2}}}[\boldsymbol{F}\boldsymbol{Z} - \boldsymbol{F}\boldsymbol{X}\boldsymbol{\beta}]\right\}_n \geq \Phi^{-1}\left(\frac{\alpha}{2}\right), \quad e.x. \ \alpha = 0.05$$

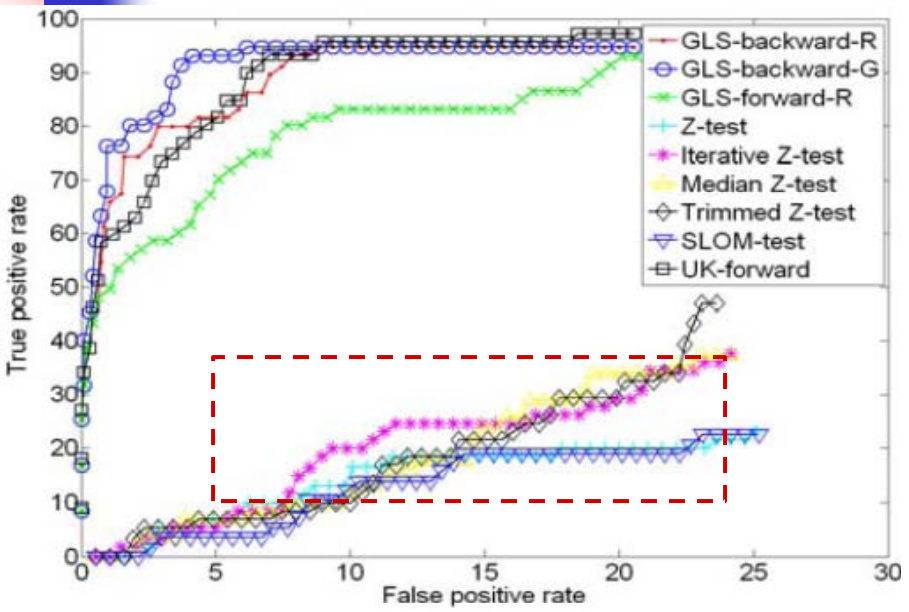
Connections with Existing Methods

- If $F\sum F^T = \sigma^2 I$, then *GLS-SOD* is equivalent to *Universal Kriging SOD*.
  - Local vs. global estimator for a spatial Gaussian random field. The key is which estimator is more robust.
- When $F\sum F^T = \sigma^2 I$, $FXB = \mu I$ and $FF = \sigma_0^2 I$, then *Local based SOD* is equivalent to *GLS-SOD* and *Universal Kriging SOD*.
  - *Local based SOD* is a special case of *GLS-SOD*

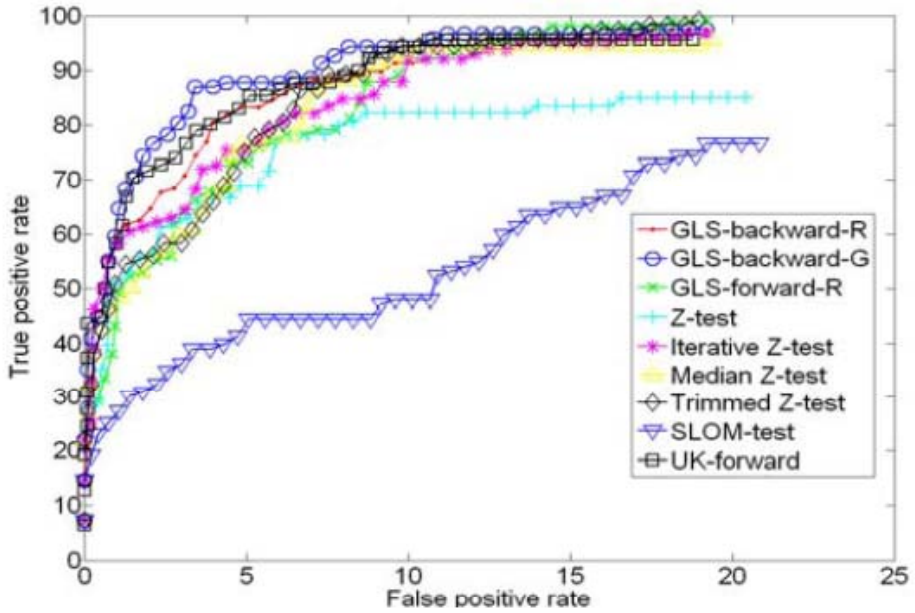
Experiments

- Standard Simulation Model for SOD
  - 864 different simulation settings. Six repetitions for each setting; consider average error
  - Existing statistical SOD methods only consider 10 to 15 simulation settings in their experiments.

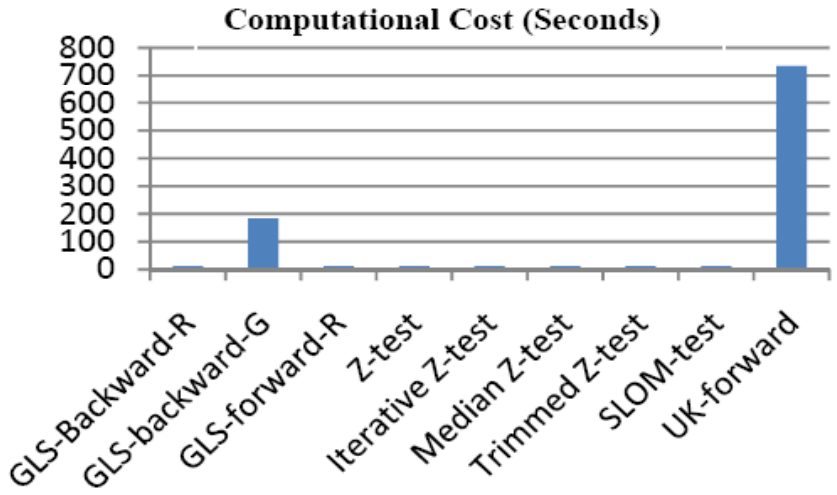
Simulation Results



f) Nonlinear trend, clustered outliers, $\alpha = 0.15, \sigma_0^2 = 10, c = 5, K$



a) Constant trend, isolated outliers, $\alpha = 0.1, \sigma_0^2 = 2, c = 15, K = 4$



Summary

- Design of a generalized local statistical framework
- Robust estimation and outlier detection methods based on the proposed GLS framework
- In-depth study on the connection between different SOD methods
- Comprehensive simulations to validate the effectiveness and efficiency of GLS

Thank you !

chenf@vt.edu

Property of "F∑F^T" (Theorem 2)

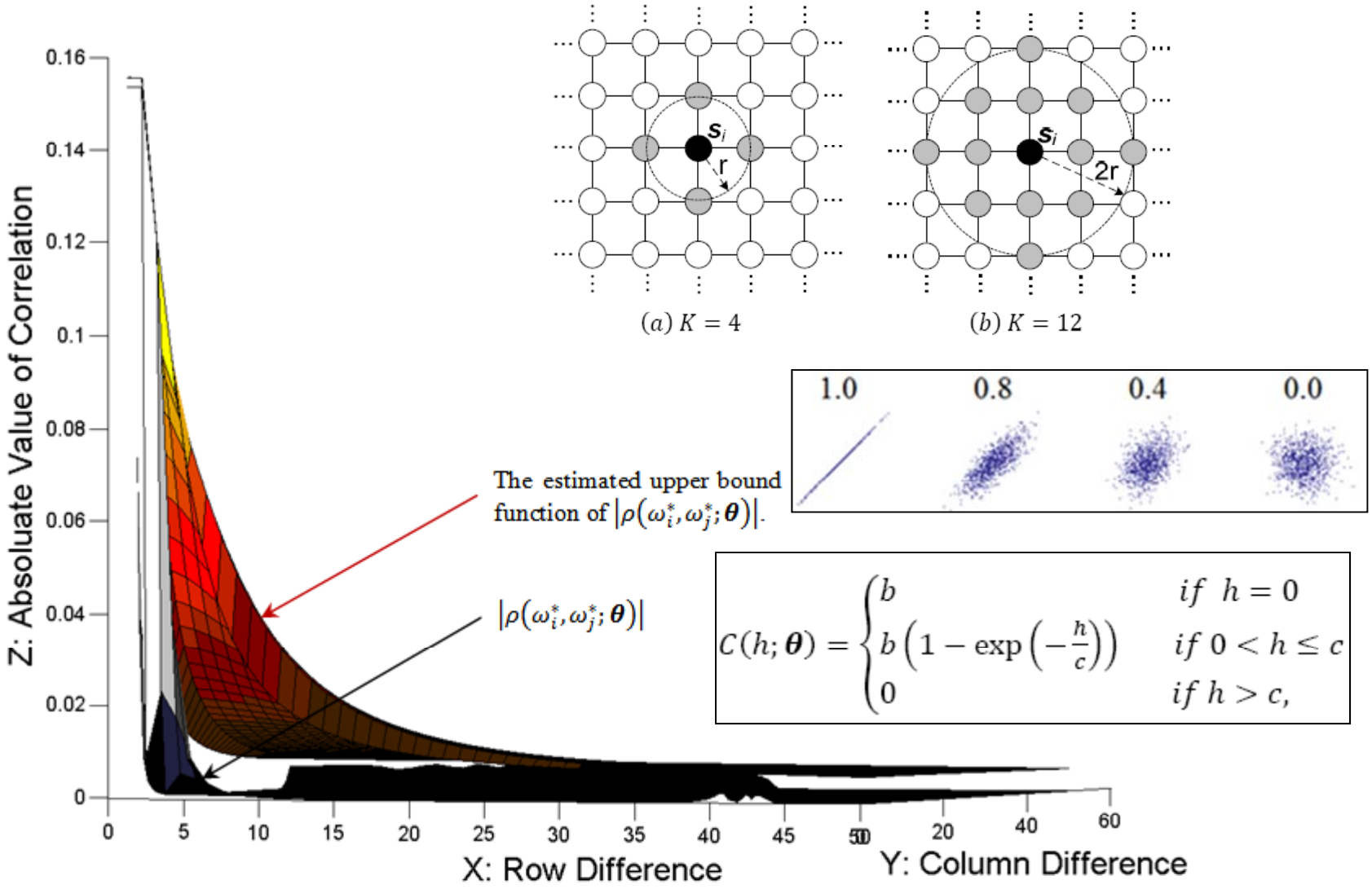


Figure 9: The comparison between the true correlation $|\rho(\omega_i^*, \omega_j^*; \boldsymbol{\theta})|$ and the estimated bound function. Here, $K = 12, c = 40$.

Property of "$\sigma_0^2 FF^T$"

Theorem 1: *The random vector* $\boldsymbol{e}^* = \boldsymbol{Fe}$ *has two major properties*

1) *The variance* $Var(e_i^*) = \frac{K+1}{K}\sigma_0^2, i = 1 \dots n,$

2) *The correlation* $\left|\rho\left(e_i^*, e_j^*\right)\right| \leq \frac{2}{K+1}, \forall i, j \text{ with } i \neq j,$

where e_i^* *refers to the i-th element in the vector* $\boldsymbol{e}^*$.

Conclusion: When selecting a relatively large neighborhood size to do local smoothing, we can approximate "FF^T" as an identity matrix.

$$\mathbf{diff}(\boldsymbol{Z}) \sim \mathbf{N}(\boldsymbol{FX\beta}, \boldsymbol{F\Sigma F}^T + \boxed{\sigma_0^2 \boldsymbol{FF}^T})$$

Connections with Existing Methods

Theorem 5: *Suppose that* $\boldsymbol{F\Sigma F^T = \sigma^2 I}$ *and the parameters of Kriging-SOD and GLS-SOD are correctly calculated by robust estimation, then Kriging-SOD and GLS-SOD are equivalent.*

Theorem 6. *If* $\boldsymbol{F\Sigma F^T = \sigma^2 I}$, $\boldsymbol{\sigma_0^2 FF^T = \sigma_0^2 I}$, *the parameters of GLS-SOD and LS-SOD are correctly calculated by robust estimation, and one of the following conditions is true, then GLS-SOD becomes equivalent to LS-SOD.*

(1) $\boldsymbol{Z(s)}$ *has a constant trend (mean):* $\boldsymbol{X\beta} = \mathbf{cI}$, *where c is a constant value.*

(2) $\boldsymbol{Z(s)}$ *is a linear trend of spatial coordinates, and each point* $\boldsymbol{s}$ *is the geometric center (or centroid) of its neighbors.*

Simulations

- Simulation Model and Settings
  - 864 different simulation settings. Six repetitions for each setting; consider average error

$$\boldsymbol{Z} = \boldsymbol{X\beta} + \boldsymbol{\omega} + \boldsymbol{e} \sim N(\boldsymbol{X\beta}, \boldsymbol{\Sigma} + \sigma_0^2 \boldsymbol{I})$$

$$\boldsymbol{\omega} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$$

$$C(h;\boldsymbol{\theta}) = \begin{cases} b & if\ h = 0 \\ b\left(1 - \exp\left(-\frac{h}{c}\right)\right) & if\ 0 < h \le c \\ 0 & if\ h > c, \end{cases}$$

Exponential Model

$$\epsilon(\boldsymbol{s}) \sim \begin{matrix} N(0, \sigma_0^2) & \text{with probability } 1-\alpha \\ N(0, \sigma_C^2) & \text{with probability } \alpha \end{matrix}$$

Contaminations

$$C(h;\boldsymbol{\theta}) = \begin{cases} b & \\ b\left(1 - \frac{3h}{2c} + \frac{1}{2}\left(\frac{h}{c}\right)^3\right) & \\ 0 & \end{cases}$$

Spherical Model

$$\boldsymbol{x}(\boldsymbol{s}) = [1, x(\boldsymbol{s}), y(\boldsymbol{s}), x(\boldsymbol{s}) \cdot y(\boldsymbol{s}), x(\boldsymbol{s})^2, y(\boldsymbol{s})^2\ ]$$

Simulation Results

Table 3: Competition statistics for different combinations of parameter settings. Each cell contains three values, representing the win times for the related method based on the accuracies of top 10, 15, and 20 ranked outlier candidates for all methods.

Algorithm	Constant Trend	Linear Trend	Nonlinear Trend
GLS-backward-R	47, 47, 45	79, 72, 82	76, 81, 77
GLS-backward-G	88, 86, 89	114, 102, 120	141,144, 138
GLS-forward-R	13, 11, 14	22, 25, 27	40, 36, 47
Z-test	47, 35, 40	29, 30, 13	0, 0, 0
Iterative Z-test	35, 46, 63	16, 20, 21	0, 0, 0
Median Z-test	20, 23, 29	1, 7, 8	0, 0, 0
Trimmed Z-test	15, 23, 32	5, 13, 13	0, 0, 0
SLOM-test	0,0, 0	0, 0, 0	0, 0, 0