2012 IEEE 12th International Conference on Data Mining

Robust Prediction and Outlier Detection for Spatial Datasets

Xutong Liu, Feng Chen, Chang-Tien Lu
Department of Computer Science, Virginia Tech
7054 Haycock Road, Falls Church, VA, 22043
{xutongl, chenf, ctlu}@vt.edu

***Abstract*—Spatial kriging is a widely used predictive model for spatial datasets. In spatial kriging model, the observations are assumed to be Gaussian for computational convenience. However, its predictive accuracy could be significantly compromised if the observations are contaminated by outliers. This deficiency can be systematically addressed by increasing the robustness of spatial kriging model using heavy tailed distributions, such as the Huber, Laplace, and Student's t distributions. This paper presents a novel Robust and Reduced Rank Spatial Kriging Model (R^3-SKM), which is resilient to the influences of outliers and allows for fast spatial inference. Furthermore, three effective and efficient algorithms are proposed based on R^3-SKM framework that can perform robust parameter estimation, spatial prediction, and spatial outlier detection with a linear-order time complexity. Extensive experiments on both simulated and real data sets demonstrated the robustness and efficiency of our proposed techniques.**

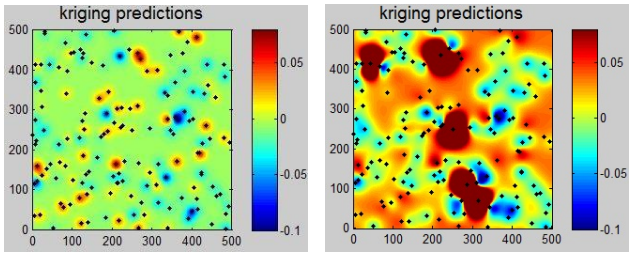
***Keywords*-Robust Estimation; Laplace Approximation; Outlier Detection;**

I. Introduction

With the increasing public sensitivity and concern on environmental issues, as well as the development of remote sensing technologies, huge amounts of spatial data have been collected from location based social network applications to scientific data, and the volume keeps increasing at fast pace over recent decades. As one of the major research issues, the prediction of spatial data has attracted significant considerations. Illustrative applications include climate prediction, environmental monitoring, molecular dynamical pattern mining, and infectious disease outbreak prediction.

Spatial prediction is the process of estimating the values of a target quantity at unobserved locations. Given the large volume of spatial data, it is computationally challenging to apply traditional prediction methods in either an allowable memory space limit or an acceptable time limit, even in supercomputing environments. Efficient prediction for large spatial data has therefore become one of the emerging challenges in data mining fields. Most existing spatial prediction methods have the time complexity of $O(n^3)$. Recently, a number of approximate methods have been proposed to tackle the "Big N" problem using different techniques, such as kernel convolutions [1], low rank basis functions or splines [2], moving averages, likelihood approximation [3], and Markov random field [4]. Recent advance by Banerjee et al. [5] proposed a reduce rank spatial kriging approach that projects the spatial process onto a subspace generated by realizations of the original process at a specific set of locations named as knots. All these methods assume that the observations follow a multivariate Gaussian distribution.

However, a well-known limitation with the above Gaussian observation model is non-robustness. The estimations of the mean and variance-covariance matrices are sensitive to outliers due to the well-known masking and swamping effects [6]. In addition to the impacts on parameters estimation, outliers also significantly reduce the accuracy of spatial predictions. For example, a single corrupted observation will deviate the posterior expectation of predictions at unobserved locations far away from the level described by the other observations. As demonstrated by Figure 1, the kriging prediction result is heavily distorted by the existence of 5 outliers (in the dark red areas in Figure 1(b)). This limitation can be properly addressed by considering robust statistics techniques.



(a) Original Data (b) Contaminated Data

Figure 1: Impacts of spatial outliers on prediction

Currently, a number of robust methods have been proposed for different learning problems, including multivariate regression, Kalman filtering and smoothing, clustering, and independent component analysis. The majority of these methods can be summarized by using a probabilistic framework [6] in which the measurement error is modeled by a heavy tailed distribution, such as the Huber, Laplace, Student's t, and Cauchy distributions, instead of the traditional Gaussian distribution. The prediction problem can then be reformulated as a Maximum-A-Posterior (MAP) prediction problem conditional on observations. However, employing heavy tailed distributions makes the prediction process analytically intractable. Although stochastic simulation methods

1550-4786/12 $26.00 © 2012 IEEE
DOI 10.1109/ICDM.2012.147

have been applied to estimate an approximate posterior distribution, for example, via MCMC or particle filtering [7] , these versatile methods are very computationally intensive. Jylanki et al. [8] presented an efficient expectation propagation algorithm for robust Gaussian process regression based on the Student's t distribution, while Svensn and Bishop [9] proposed a variational inference approach to robust Student's t mixture clustering. Gandhi and Mili [10] proposed a robust Kalman filter based on the Huber distribution and the iterative reweighted least squares (IRLS) method. An efficient Kalman smoother was presented by Aravkin et al. [11] based on the Laplace distribution and the convex composite extension of the Gauss-Newton method.

This paper aims to address the robust prediction problem for large spatial dataset. It considers the same probabilistic framework as that was used in existing robust methods. Specifically, a Robust and Reduced-Rank Spatial Kriging Model (R^3-SKM) is formulated, and then efficient algorithms are proposed by utilizing Laplace approximation to perform parameter estimation, robust spatial prediction, and spatial outlier detection.

To the best of our knowledge, this is the first statistical approach that can perform robust spatial prediction in linear time. The main contributions can be summarized as follows:

- **Formulation of the R^3-SKM model.** A Robust and Reduced Rank Spatial Kriging Model is proposed in which the measurement error is modeled by a heavy tailed distribution, and a Bayesian hierarchical framework is integrated to support priors on model parameters.
- **Design of an approximate algorithm for robust parameter estimation.** The posterior distribution of latent variables conditional on parameters and observations is estimated via Gaussian approximation. Furthermore, the posterior distribution of parameters conditional on observations is estimated via Laplace approximation. It has time complexity of $O(n)$.
- **Development of robust inference algorithms.** R^3-SP (Robust and Reduced Rank Spatial Prediction) and R^3-SOD (Robust and Reduced Rank Spatial Outlier Detection) algorithms are proposed to perform robust spatial prediction and spatial outlier detection. Their time complexities are analyzed, which scale linearly.
- **Comprehensive experiments to validate the robustness and efficiency of the proposed techniques**. The R^3-SKM was evaluated by the extensive experiments on simulated and real datasets. The results demonstrated that the three algorithms based on R^3-SKM outperformed existing representative techniques, when the data were contaminated by outliers.

The rest of paper is organized as follows. Section II reviews the theoretical background. Section III presents the R^3-SKM framework. A general approach based on R^3-SKM is proposed to perform robust parameter estimation in Section IV, and two inference algorithms are discussed in Section V. Experiments on both simulated and real datasets are presented in Section VI. The paper concludes with a summary of the research in Section VII.

II. Theoretical Background

This section reviews the Spatial Kriging Model (SKM) and knot based reduced-rank techniques.

A. Spatial Kriging Model

Let us define a numerical random field $Y(s)$ on a domain $D \subseteq \mathcal{R}^2$, and $Y = (Y(s_1), \cdots, Y(s_n))'$ be the $n \times 1$ vector of observed responses, each of which is along with a $p \times 1$ vector of spatially referenced predictors $x(s)$. The associated spatial kriging model can be represented as

$$Y(s) = x^T(s)\beta + \eta(s) + \epsilon(s) \tag{1}$$

where $\epsilon(s)$ is a spatial white noise process with mean zero, $var(\epsilon(s)) = \tau^2 > 0$, and τ^2 is a parameter to be estimated. The white noise assumption implies that $R_{i,j}(\phi) = cov(\epsilon(s_i), \epsilon(s_j)) = 0$, unless $i = j$. $x(s)$ refers to a vector of known predictors, and the coefficients β are unknown. $x^T(s)\beta$ is a vector of deterministic (spatial mean) or trend functions, which models large scale variations, and the spatial random process $\eta(s)$ captures the small scale variations. The hidden process $\eta(s)$ captures spatial association. It is assumed to follow a Gaussian process with zero mean and the covariance function $\sigma^2 C(s, s'; \phi)$, where σ^2 refers to the variance, and $C(\cdot; \phi)$ the correlation function of the process controlled by the parameter ϕ. Function C controls the smoothness and scale among latent variables $\eta(s_i)$, and can be selected freely as long as the resulting covariance matrix is symmetric and positive semi-definite.

B. Reduced Rank Methodology

The spatial inference (e.g., spatial prediction, outlier detection) based on the SKM model involves the inversion of the n by n correlation matrix, which has the time complexity of $O(n^3)$. This makes the SKM model prohibitively expensive for large n. The knot-based model proposed by Banerjee et al. [12] considers a fixed set of "knots" $S^* = (s_1^*, \cdots, s_{n^*}^*)$ with $n^* \ll n$. The Gaussian process $\eta^*(s)$ yields an n^*-vector of realizations over the knots, that is, $\eta^* = (\eta(s_1^*), \cdots, \eta(s_{n^*}^*))$, which follows a $GP\{0, C^*(s_i^*, s_j^*; \theta)\}$. Spatial estimation at a generic site s is operated through

$$\tilde{\eta}(s) = E\{\eta(s)|\eta^*\} = c^T(s; \theta)C^{*-1}(\theta)\eta^* \tag{2}$$

where $c(s; \theta) = [C(s, s_j^*; \theta)_{j=1}^{n^*}]$. The reduced rank SKM model can be formalized as

$$Y(s) = x^T(s)\beta + \tilde{\eta}(s) + \epsilon(s) \tag{3}$$

Table I: Description of Major Symbols

Sym.	Description
S	$S = \{s_i\}_{i=1}^n$, a set of n training locations.
S^*	$S^* = \{s_i^*\}_{i=1}^m$, a set of m knot locations.
Y	A given set of observations with numerical attributes which follow Gaussian distribution. $Y = \{Y(s_i)\}_{i=1}^n$
X	A set of explain variables. $\{X(s_i)\}_{i=1}^n$ is a $p \times 1$ vector of covariates or explain variables at location s_i.
η	Spatial random effects of the observations, which provide local adjustments to the means. $\eta = \{\eta(s_i)\}_{i=1}^n$
η^*	Spatial random effects of the knots. $\eta^* = \{\eta^*(s_i)\}_{i=1}^m$
$\tilde{\eta}$	The predicted values of η by η^*.$\tilde{\eta} = \{\tilde{\eta}(s_i)\}_{i=1}^n$
$\tilde{\epsilon}$	$\{\tilde{\epsilon}(s_i)\}_{i=1}^n$ is the nugget measurement error.
v^*	$v^* = (\eta^*, \beta')'$, it is a $(m+p) \times 1$ vector comprising the realizations of the spatial predictive process and the regression parameters.
H	$H = [F(\phi)X]$. $F(\phi)$ is a transformation matrix which describes that $\tilde{\eta}$ is defined as a spatially varying linear transformation of η^*.
Θ	The set of sample locations of θ, based on the mode and Hessian at it of $\hat{\pi}(\theta\|Y, Z)$. $\Theta = \{\theta\}_{k=1}^K$.
Δ	The set of weight values of sample θ, which are computed by their corresponding posterior distributions. $\Delta = \{\Delta\}_{k=1}^K$

It is important to select a reasonable number of knots as well as their spatial locations. This is related to the problem of spatial design. There are two popular knots selection strategies. One is to draw a uniform grid to cover the study region and each grid is considered as a knot. Another is to place knots such that each covers a local domain and the regions with dense data have more knots. In practice, it is feasible to validate models by using different number of knots and different choices of knots to obtain a reliable and robust configuration.

III. Robust and Reduced Rank Spatial Kriging Model

The Robust and Reduced-Rank Spatial Kriging Model (R^3-SKM) integrates robust, reduced-rank, and Bayesian hierarchical techniques together.

The proposed R^3-SKM is defined as

$$Y = X\beta + \tilde{\eta} + \tilde{\epsilon} \tag{4}$$

in which most of the variables are defined in section II, except the measurement error $\tilde{\epsilon}$ now follows a heavy tailed distribution with the probability density function $f(\tilde{\epsilon}; \mu, \varrho^2) = \frac{1}{\varrho} h((\tilde{\epsilon} - \mu)/\varrho)$, where μ refers to the mean, and ϱ the dispersion parameter. Examples of the h function include: 1) Laplace distribution: $h(x) = \frac{1}{2} e^{-|x|}$; 2) Student's t distribution: $h(x) = c(x + \nu)^{(p+\nu)/2}$, where c is a normalization constant, the case $v = 1$ is the Cauchy density, and the limiting case $v \to \infty$ yields the normal distribution; and 3) Huber distribution: $h(x) = ce^{-\varphi(x;\varrho)}$,

$$\varphi(x;\kappa) = \begin{cases} \kappa|x| - \frac{1}{2}\kappa^2, & \text{for } |x| > \kappa \\ \frac{1}{2}x^2, & \text{for } |x| \le \kappa, \end{cases} \tag{5}$$

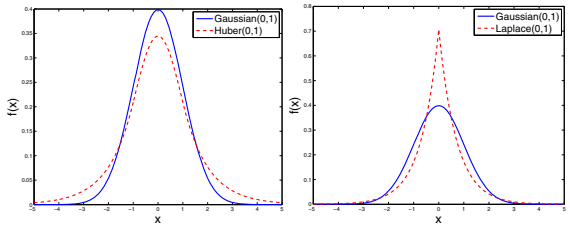


Figure 2: pdfs of Heavy Tailed Distributions

where c is a normalization constant that ensures $\int \frac{c}{\varrho} e^{-\varphi(x;\kappa)} = 1$, and ϱ is a range parameter of the distribution. The probability density functions (pdf) of the Huber, Laplace and Gaussian distribution are compared in Figure 2. The robustness of the R^3-SKM is realized by the latent variation component $\tilde{\epsilon}_i, i = \{1, \cdots, n\}$, which follows a heavy tailed distribution. For example, the parameter ν in Student's t, or κ in Huber distribution controls the degree of the robustness. When the value of ν increases, the robustness of the Student's t will decrease.

R^3-SKM can be formalized in the framework of Bayesian hierarchical model with three layers, including the observation layer, the latent robust Gaussian process layer, and the parameter layer. The observation layer contains the observations $Y = \{Y(s_1), \cdots, Y(s_n)\}$. It is assumed that each $Y(s_i)$ follows a Gaussian distribution. Each random variable $Y(s_i)$ is related to the latent Gaussian effects in the second layer, $v^* = (\eta^{*\prime}, \beta')'$, which is the $(m+p) \times 1$ vector. Specifically, β is assigned a multivariate Gaussian prior, i.e., $\beta \sim \mathcal{N}(\mu_\beta, \Sigma_\beta)$. The third level of the hierarchical model consists of the related parameters with the latent variables. In the R^3-SKM model, the parameters include $\theta = (\sigma^2, \phi, \nu, \varrho^2)$. That is, σ^2 and ϕ for modeling η^*, μ and ϱ^2 for modeling $\tilde{\epsilon}$. σ^2 has an inverse gamma prior distribution: $\sigma^2 \sim IG(\alpha_\sigma, \gamma_\sigma)$, where α_σ and γ_σ are sufficiently small informative prior distribution. The correlation parameter ϕ is usually assigned an informative prior decided based on the underlying spatial domain, i.e., $\phi \sim \mathcal{U}(a_\phi, b_\phi)$, a uniform distribution over a finite range. In Student's t distribution, $\nu \sim \mathcal{U}(a_\nu, b_\nu)$ and $\varrho^2 \sim IG(\alpha_\varrho, \gamma_\varrho)$. Taking the Student's t as an example, the graphic representation of the R^3-SKM is depicted in Figure 3.

IV. Robust Parameter Estimation

This section presents a novel approach, R^3-PE (Robust and Reduced-Rank Parameter Estimation), to execute the robust parameter estimation by integrating Laplace approximation [13]. It consists of two critical steps: 1) Gaussian approximation of the posterior distribution of latent variables conditional on parameters and observations; 2) Laplace approximation of the posterior distribution of corresponding parameters conditional on observations. Student's t distribution is selected to model the *pdf* of $\tilde{\epsilon}$.

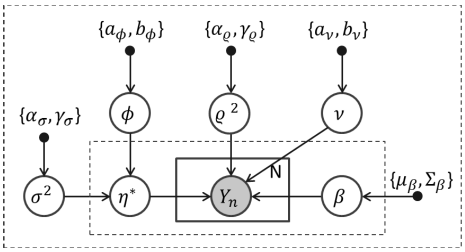


Figure 3: Graphic Model Representation of R^3-SKM

A. Gaussian Approximation of Posterior Distribution of v^*

First, we need to compute $\pi(v^*|Y,\theta)$, where $v^* = (\eta^{*'}, \beta')'$, which consists of the spatial predictive process and the regression parameters. Its mean and covariance matrix are computed as follows.

$$\mu_{v^*} = (0_m, \mu_\beta)', \Sigma_{v^*} = \begin{bmatrix} \sigma^2 C^*(\phi) & 0_{m\times p} \\ 0_{p\times m} & \Sigma_\beta \end{bmatrix} \tag{6}$$

we have prior $v^* \sim N(\mu_{v^*}, \Sigma_{v^*})$.

With the information depicted by the graphical model in Figure 3, we determine that the full conditional distribution of $\pi(Y|v^*,\theta)$ follows the heavy tailed distribution, which can be approximated as a Gaussian distribution of v^* by Taylor expansion. For example, if $\tilde{\epsilon}$ accords to the Student's t distribution, then $\pi(y_i|\nu, T_iHv^*, \varrho^2) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu\varrho^2}}(1 + \frac{1}{\nu}\frac{(y_i - T_iHv^*)^2}{\varrho^2})^{-\frac{\nu+1}{2}}$, where $T_iHv^* = x_i\beta + \tilde{\eta}_i$ and T_i is the i^{th} row of unit matrix I_n. H=$[F(\phi)X]$, and $F(\phi) = \mathcal{C}(\phi)'C^{*-1}(\phi)$, where $\mathcal{C}(\phi)'$ is an $n \times m$ matrix whose i^{th} row is the $1 \times m$ vector in which the j^{th} element is given by $C(s_i, s_j; \phi)$. We Taylor expand it to second order by expressing the result in a quadratic form of v^*,

$$\begin{aligned}
&\log(\hat{\pi}(Y|v^*,\theta)) = -\frac{1}{2}v^{*'}Q_Y v^* + v^{*'}b_Y + const \\
&Q_Y = \sum_{i=1}^{n} Q_{y_i}, b_Y = \sum_{i=1}^{n} b_{y_i} \\
&Q_{y_i} = m_1 m_2 \left(\frac{m_2 \nabla D(\hat{v}^*)\{\nabla D(\hat{v}^*)\}'}{(1 + m_2 D(\hat{v}^*))^2} - \frac{\nabla^2 D(\hat{v}^*)}{1 + m_2 D(\hat{v}^*)}\right) \\
&b_{y_i} = \frac{m_1 m_2 \nabla D(\hat{v}^*)}{1 + m_2 D(\hat{v}^*)} + Q_{y_i}\hat{v}^* \\
&m_1 = -\frac{\nu+1}{2}, m_2 = \frac{1}{\nu\varrho^2} \\
&D(\hat{v}^*) = (y_i - T_iH\hat{v}^*)'(y_i - T_iH\hat{v}^*) \\
&\nabla D(\hat{v}^*) = -2H'T_i'y_i + 2H'T_i'T_iH\hat{v}^* \\
&\nabla^2 D(\hat{v}^*) = 2H'T_i'T_iH
\end{aligned} \tag{7}$$

With the above Gaussian approximation, $\pi(v^*|Y,\theta)$ is analytically available and numerical routines can be applied. For the R^3-SKM, the full conditional for v^* is

$$\begin{aligned}
\pi(v^*|Y,\theta) &\propto \hat{\pi}(Y|v^*,\theta)\pi(v^*|\theta) \\
&\propto exp[-\frac{1}{2}v^{*'}Qv^* + v^{*'}b]
\end{aligned} \tag{8}$$

where the full conditional precision matrix $Q = Q_Y + \Sigma_{v^*}^{-1}$, and the canonical parameter $b = b_Y + \Sigma_{v^*}^{-1}\mu_{v^*}$. Thus, the full conditional is $\pi(v^*|Y,\theta) \sim N(Q^{-1}b, Q^{-1})$. We can compute the required inverse and determinant of the size $(m+p)\times(m+p)$ matrix Q by utilizing the structure of Σ_{v^*} and H. The main cost of matrix inversion is thus $O(m^3)$, since the number of knots is m, assuming $m \gg p$.

B. Laplace Approximation of Posterior Distribution of θ

Different from $\pi(v^*|Y,\theta)$, the posterior $\pi(\theta|Y)$ is usually skewed and the approximation as a Gaussian distribution is inappropriate. The posterior $\pi(\theta|Y)$ plays an important role in the inference of marginal posterior of latent variables that are of interests. Take v^* as an example, the interest is to estimate the marginal posterior $\pi(v^*|Y)$, which has

$$\pi(v^*|Y) = \int \pi(v^*|Y,\theta)\pi(\theta|Y)d\theta \tag{9}$$

It is possible to obtain a sample set of $\{\theta_1, \cdots, \theta_K\}$ from the input space of θ that represents an approximate discrete form of the posterior $\pi(\theta|Y)$. We can estimate the approximate $\hat{\pi}(v^*|Y)$ by

$$\hat{\pi}(v^*|Y) = \sum_{k=1}^{K} \pi(v^*|Y,\theta_k)\pi(\theta_k|Y)\Delta_k \tag{10}$$

where Δ_k is the weight of the sample θ_k that can be measured by its normalized probability density. The critical step is to efficiently identify a suitable sample set $\{\theta_1, \cdots, \theta_K\}$, as well as its corresponding weight set $\{\Delta_1, \cdots, \Delta_K\}$. The posterior $\pi(\theta|Y)$ can be reformalized as

$$\pi(\theta|Y) \propto \frac{\pi(Y|v^*,\theta)\pi(v^*|\theta)\pi(\theta)}{\pi(v^*|Y,\theta)} \tag{11}$$

Laplace Approximation (LA) can be applied to approximate the denominator $\pi(v^*|Y,\theta)$ as a Gaussian distribution, and then set the vector of variables, v^*, to the mode. The LA method uses similar ideas for Bayesian spatial inference:

$$\hat{\pi}(\theta|Y) \propto \left.\frac{\pi(Y|v^*,\theta)\pi(v^*|\theta)\pi(\theta)}{\hat{\pi}(v^*|Y,\theta)}\right|_{v^*=\hat{v}^*} \tag{12}$$

where $\hat{\pi}(v^*|Y,\theta)$ is a Gaussian approximation as shown in Equation (8). We can get the mode $\hat{v}^*$ and the curvature at the mode of this full conditional expression. The preceding Gaussian approximation can be efficiently conducted by using the popular Iterated Re-weighted Least Squares (IRLS) algorithm. The above detailed procedures can be summarized as Algorithm 1.

Algorithm 1 iterates l_1 times, from Step 2 to Step 8 until convergence. Among these steps, Step 6 has the highest

Algorithm 1 Exploring posterior distribution of $\pi(\theta|Y)$

Input: S, S^*, Y, X
Output: Θ, Δ

1: Choose an initial value $\theta = (\sigma^2, \phi, \nu, \varrho)$;
2: **repeat**
3: Construct μ_{v^*}, Σ_{v^*} with θ (See Equation(6));
4: Calculate the transformation matrix H;
5: Gaussian approximation of $\pi(Y|v^*, \theta)$ as the form of v^*.
6: Apply IRLS to identify the mode $\hat{v}^*$ and Hessian at the mode of $\hat{\pi}(v^*|Y, \theta)$.
7: Compute the gradient and Hessian of $\hat{\pi}(\theta^*|Y)$ and apply one Newton's step to update θ.
8: **until** Convergence
9: Explore the contour of $\tilde{\pi}(\theta|Y)$ based on its mode and Hessian at the mode, obtain K sample locations, $\Theta = \{\theta_1, ..., \theta_K\}$.
10: Compute and normalize $\{\hat{\pi}(\theta_1|Y), \cdots, \hat{\pi}(\theta_K|Y)\}$ to obtain $\Delta = \{\Delta_1, \cdots, \Delta_K\}$ as $\Delta_k = \frac{\hat{\pi}_k(\theta_k|Y)}{\sum_{k=1}^K \hat{\pi}_k(\theta_k|Y)}$.

time cost, because the solution is analytically intractable and numerical optimization techniques are applied. An efficient IRLS algorithm is proposed to conduct this process. Step 6 is first reformulated as the following optimization problem

$$\underset{v^*}{\operatorname{argmax}} \hat{\pi}(v^*|Y, \theta) = \underset{v^*}{\operatorname{argmin}} - \ln \pi(Y|v^*, \theta) - \ln \pi(v^*|\theta) \quad (13)$$

Expanding the density functions $\pi(Y|v^*, \theta)$ and $\pi(v^*|\theta)$, we have that $\underset{v^*}{\operatorname{argmin}}\{\frac{1}{2}v^{*\prime}\{\sum_{i=1}^n Q_i\}v^* - v^{*\prime}\{\sum_{i=1}^n b_i\} + \frac{1}{2}\left(v^* - \mu_{v^*}\right)'\Sigma_{v^*}^{-1}(v^* - \mu_{v^*})\}$. The gradient and Hessian matrix of the above objective function can be obtained as

$$\begin{aligned} \bigtriangledown \hat{\pi}(v^*|Y, \theta) &= (\sum_{i=1}^n Q_i + \Sigma_{v^*}^{-1})v^* - (\sum_{i=1}^n b_i + \Sigma_{v^*}^{-1}\mu_{v^*}) \\ \bigtriangledown^2 \hat{\pi}(v^*|Y, \theta) &= \sum_{i=1}^n Q_i + \Sigma_{v^*}^{-1} \end{aligned} \quad (14)$$

The IRLS algorithm for Step 6 is described as follows:

1) Select an initial $\hat{v}^*$
2) Until convergence

Update $\hat{v}^* = \hat{v}^* - \left(\bigtriangledown^2 \hat{\pi}(\hat{v}^*|Y, \theta)\right)^{-1} \bigtriangledown \hat{\pi}(\hat{v}^*|Y, \theta)$.

3) Output $\hat{v}^*$ as the mode of $\hat{\pi}(v^*|Y, \theta)$.

Computational Complexity. In Algorithm 1, suppose that it needs l_2 iterations to find the mode $\hat{v}^*$ and Hessian at the mode of $\hat{\pi}(v^*|Y, \theta)$, the time cost of Step 6 is $O(l_2 * (n * m^2 + m^3))$. The Step 5, Gaussian approximation of $\pi(Y|v^*, \theta)$, takes $O(n * m)$. Overall, Steps 2-8, which generate the converged gradient and Hessian of $\pi(\theta|v^*)$ take $O(l_1 * l_2 * (n * m^2 + m^3) + l_1 * n * m)$. Finally, sampling the θ set and computing their corresponding weighted values take $O(K)$. In summary, assuming $n \gg K$, $n \gg m$, $n \gg l_1$ and $n \gg l_2$, the total computational complexity of robust parameter estimation based on R^3-SKM is $O(n)$.

Algorithm 2 Robust Reduced Rank Spatial Prediction

Input: $S, S^*, S^0, Y, X, X^0, \Theta, \Delta$
Output: Y^0

1: **for** $k = 1$**to**K **do**
2: Construct μ_{v^*}, Σ_{v^*} with θ_k and S^* (See Equation (6)).
3: Calculate the transformation matrix H with θ_k, S^*, S, X.
4: Gaussian approximation of the likelihood of Y.
5: Calculate the mode, Hessian at the mode of $\hat{\pi}(v^*|Y, \theta_k)$, and its Gaussian approximation (See Equation (8)).
6: Predict Y_k^0 for new locations S^0. (See Equation (15))
7: **end for**
8: Calculate the final Y^0 values as $Y^0 = \sum_{k=1}^K Y_k^0 \times \Delta_k$

V. Robust Spatial Inference

This section formalizes the Robust and Reduced Rank Spatial Prediction (R^3-SP) , and Robust and Reduced Rank Spatial Outlier Detection (R^3-SOD) based on the R^3-SKM.

A. Robust Spatial Prediction

Given a set of unsampled locations $\{s_1^0, \cdots, s_{N_{te}}^0\}$, we are interested in predicting the Y values at these locations, denoted as $Y^0 = [Y(s_1^0), \cdots, Y(s_{N_{te}}^0)]$. The first step is to estimate the posterior distributions of the corresponding latent variables $\pi(v_0|Y)$, where $v^0 = [v(s_1^0), \cdots, v(s_{N_{te}}^0)]'$. Then, the posterior distributions of Y^0 can be obtained as

$$\pi(Y^0|Y) = \int \pi(Y^0|v^0)\pi(v^0|Y)dv^0 \quad (15)$$

Given the approximated $\hat{\pi}(v^*|Y, \theta)$ and $\tilde{\pi}(\theta|Y)$ as obtained in Sections IV.A and IV.B, the posterior distribution $\pi(v^0|Y)$ can be estimated by

$$\begin{aligned} \pi(v^0|Y) &= \int\int \pi(v^0|v^*, Y, \theta)\pi(v^*|Y, \theta)\pi(\theta|Y)dv^*d\theta \\ &= \int \left\{\int \pi(v^0|v^*, \theta)\pi(v^*|Y, \theta)dv^*\right\}\pi(\theta|Y)d\theta \\ &\approx \sum_k \left\{\int \pi(v^0|v^*, \theta_k)\hat{\pi}(v^*|Y, \theta_k)dv^*\right\}\hat{\pi}(\theta_k|Y)\Delta_k \\ &\approx \sum_k \mathcal{N}(\tilde{\mu}, \tilde{\Sigma})\hat{\pi}(\theta_k|Y)\Delta_k \end{aligned} \quad (16)$$

where

$$\begin{aligned} &\Sigma^0 = Cov(v^0), \Sigma^* = Cov(v^*), \Sigma^{0*} = Cov(v^0, v^*) \\ &\tilde{\mu} = \Sigma^{0*}\Sigma^{*-1}Q^{-1}b \\ &\tilde{\Sigma} = \Sigma^0 - \Sigma^{0*}\Sigma^{*-1}\Sigma^{0*\prime} + \Sigma^{0*}\Sigma^{*-1}Q^{-1}\Sigma^{*-1}\Sigma^{0*\prime} \end{aligned}$$

Based on the above theoretical analysis, the main procedures of R^3-SP are described by Algorithm 2. In the R^3-SP algorithm, we first derive the K samples of θ and their weight values, Δ, by utilizing the R^3-SKM framework, and then use each generated sample, θ_k, to construct the corresponding mean and covariance matrix of latent variables, v^*. Next, the transformation matrix $H = [F(\phi)X]$ is computed, in which $F(\phi)$ describes the spatially varying

linear transformation of $\tilde{\eta}$ on η^*. Furthermore, the likelihood of Y are approximated as the result of a quadratic form of v^*. Next, the mode of $\hat{\pi}(v^*|Y, \theta_k)$ are calculated to predict the new observations Y_k^0 at sample θ_k. Finally, the predicted Y is calculated as $Y^0 = \sum_{k=1}^{K} Y_k^0 \times \Delta_k$.

Computational complexity. Similarly, for the R^3-SP algorithm, Steps 4 and 6 dominate most time costs, because they are naturally analytical intractable. With the numerical optimization discussed in Section IV, it takes $O(n * m)$ to operate a Gaussian approximation of Y at each sample θ_k. And computing the mode and Hessian of $\hat{\pi}(v^*|Y, \theta_k)$ costs $O(l_2 * (m^3 + n * m^2))$. Repeating Steps 2-6 at K sample θs takes $O(K * (n * m + l_2 * (m^3 + n * m^2)))$. In summary, the total computational complexity of the R^3-SP algorithm is $O(n)$, assuming $n \gg K$, $n \gg l_2$, $n \gg p$ and $n \gg m$.

B. Robust Spatial Outlier Detection

Statistically, spatial outlier can be interpreted as observations that have abnormally low correlations with their spatial neighbors, considering normal deviations caused by measurement error (white noise). For the regular SKM framework, when a data set contains outliers, the additional variation due to those outliers will be captured by distorting spatial dependence. The white noise component is unable to handle large deviations due to the light tailed feature of the Gaussian distribution. In comparison, the proposed R^3-SKM uses heavy tailed distribution to model the measurement error. When outliers appear, our model directly captures the additional large variation due to outliers as the measurement error, which will control the resulting accurate parameter estimation and spatial outlier detection.

Therefore, the R^3-SKM can also be utilized to identify spatial outliers as objects with higher predicted $\tilde{\epsilon}$ values(measurement error). First, we apply the R^3-SKM to accurately estimate the latent variables and parameters for the contaminated spatial dataset. Second, the estimated values are utilized to operate a spatial prediction for each observed location. Finally, the differences between observed and predicted values are computed to measure their outlying degrees. The objects which have higher measurement errors are labeled as spatial outliers.

The main procedures of spatial outlier detection are described in Algorithm 3. In the R^3-SOD Algorithm, we use the K samples of θ to predict the corresponding $\{Y_i\}_k^p (k = 1, \cdots, K)$, and the predicted $\{Y_i\}^p$ is finalized by the sum over values derived from different θ_k with weight Δ_k. If the predicted $\{Y_i\}^p$ has a large deviation compared with its original value, and this deviation is higher than the cut-off value, $c \cdot \varrho (c = 3)$, then the corresponding objects will be identified as spatial outliers.

Computational complexity. As analyzed in Algorithm 2, predicting $\{\{Y_i\}_k^p\}_{i=1}^n$ takes around $O(K(l_2 n + m^3))$. Finalizing the predicted $\{\{Y_i\}^p\}_{i=1}^n$ costs $O(n)$. In summary,

Algorithm 3 Robust Reduced Rank Spatial Outlier Detection (R^3-SOD)

Input: $S, S^*, Y, X, \Theta, \Delta$
Output: Y^0

1: Repeat Steps 1-7 in Algorithm 2 to predict $\{Y_i\}_k^p$ for each locations s_i.($k = 1, \cdots, K$, and $i = 1, \cdots, n$).
2: **for** $i = 1$**to**n **do**
3: Calculate the final $\{Y_i\}^p$ values as $\{Y_i\}^p = \sum_{k=1}^{K} \{Y_i\}_k^p \times \Delta_k$.
4: Calculate the abstract difference Diff$_i$= $|\{Y_i\}^p - Y_i|$.
5: **end for**
6: Rank the objects by sorting Diff with an descending order.
7: Label the top ones that have Diff$\geq c \cdot \varrho$ as spatial outliers.

R^3-SOD algorithm takes $O(n)$, assuming $n \gg K$, $n \gg l_2$, $n \gg p$ and $n \gg m$.

VI. Experiment

This section evaluates the robustness and efficiency of the proposed R^3-SKM model based on an analysis of simulated and real data sets. Student's t distribution was selected to model the probability density function of $\tilde{\epsilon}$. All experiments were conducted on a PC with Intel(R) Core(TM) I5-2400, CPU 3.1 Ghz, and 8.00 GB memory.

A. Experiment Setting

Dataset Description

Simulation Dataset. The simulations were generated based on the following statistical model:

$$Y(s) \quad \sim \quad \mathcal{N}(x^T\beta + \eta(s), \tau^2) \tag{17}$$

where $\eta(s)$ is from a latent spatial Gaussian process with the variogram model $Var(\eta(s_i), \eta(s_j)) = \sigma^2 C(h|\phi)$, and $h = |s_i - s_j|$. $C(h|\phi)$ refers to the spatial correlation, where ϕ is the range parameter that controls its degree. The popular exponential function was used to model $C(h|\phi)$. The parameter settings used in our experiments are shown in Table II. We also evaluated different combinations of parameters, and observed similar patterns.

Table II: Parameter settings in the simulations

Variable	Setting Description
$[N_{tr}, N_{te}]$	Training and testing points were randomly generated at N_{tr} spatial locations $\{s_i\}_{i=1}^{N_{tr}}$ and N_{te} spatial locations $\{s_i\}_{i=1}^{N_{te}}$, respectively, in the range [0,50]×[0,50] units. $N_{tr} = 300, 500$, $N_{te} = 30, 100$
β	The regression coefficient $\beta = [0.5, 1.5]'$.
σ	$\sigma^2 = 4$ in all simulations
ϕ	$\phi = 25$.
τ	The nugget variance, τ^2, was set to 0.1.
$C(h\|\phi)$	An exponential spatial correlation function $C(h\|\phi) = \sigma^2 exp(-\frac{h}{\phi})$ was used in all simulations.

Real Dataset. We validated our approach on five real datasets, namely, *Lake*, *MLST*, *BEF*, *HR*, and *House*. *Lake* was originally published by Varin et al. [14] and was used

to model trout abundance in Norwegian lakes as a function of lake acidity. *MLST* [15] came from multiple listings containing structural descriptors of houses, their sale prices, and their addresses for Baltimore, Maryland, in 1978. *BEF* [16] is a forest inventory dataset from the U.S. Department of Agriculture Forest Service, Barlett, NH. *HR* is a Boston Housing dataset from 1978 which discusses issues related to the demand for clean air. *House* contains information collected for a range of variables for all the block groups in California from the 1990 Census. Both *House* and *HR* data are included in the spBayes R package [17]. Table III summarizes the main information types of each of these datasets used in our experiments.

Table III: Settings in 5 real datasets

Dataset	Size	N_{tr}	N_{te}	Y	Y-SD
BEF	437	337	100	BE basal area	0.17
Lake	371	271	100	Trout abundance	0.007
MLST	211	150	61	House price	0.17
HR	506	406	100	House price	0.10
House	20,640	1000	200	House price	0.25

Spatial Inference Method

Spatial Estimation and Prediction Methods. There are currently two popular methods used for parameter estimation and spatial prediction. The Spatial Kriging Model(SKM) predicts unobserved values as a linear combination of the known values of observed locations, while Linear Regression[18] models data using linear predictor functions, and estimates unknown model parameters from the data.

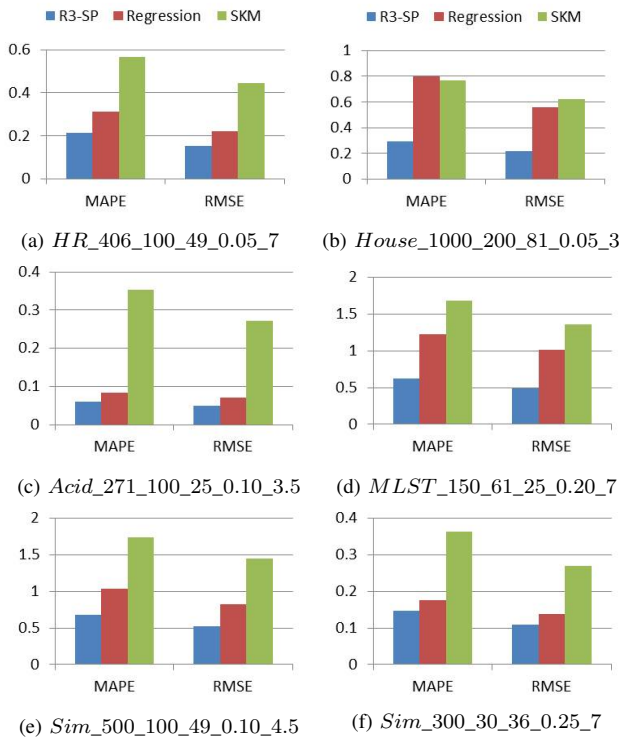
Outlier Detection Methods. We compared R^3-SOD with eight existing representative SOD approaches: Z-test [19], Median Z-test, Iterative Z-test, trimmed Z-test [20], Scatterplot [21], MoranScatterplot [22], SLOM [23] and POD [24]. The implementations of the above methods were all based on their published algorithm descriptions.

Performance Metric

Data Contamination. For each dataset, including both the simulations and real datasets, we randomly selected $\alpha\%$(contamination rate) of the data to be anomalies by shifting them from their original values with γ(shift rate) times standard deviation of Y. For each α, the synthetic outliers were generated 10 times, and the mean values of the results from the parameter estimation, spatial prediction and spatial outlier detection were calculated for each approach.

Parameter Estimation. Parameter estimation was executed only in simulations since the true values of parameters were known. We compared the estimation results from SKM, Regression and R^3-SKM, with the true values to validate their effectiveness.

Spatial Prediction. SKM, Regression and R^3-SKM were also applied in both simulation and real datasets to obtain the predicted $\tilde{Y}$. The Mean Absolute Percentage Error ($MAPE = \frac{\sum_{i=1}^{N_{te}} |Y_i - \tilde{Y}_i|}{N_{te}}$) and Root Mean Square Error



(a) $HR\_406\_100\_49\_0.05\_7$ (b) $House\_1000\_200\_81\_0.05\_3$

(c) $Acid\_271\_100\_25\_0.10\_3.5$ (d) $MLST\_150\_61\_25\_0.20\_7$

(e) $Sim\_500\_100\_49\_0.10\_4.5$ (f) $Sim\_300\_30\_36\_0.25\_7$

Figure 4: Comparison of prediction performances on simulation and real datasets

($RMSE = \{\frac{\sum_{i=1}^{N_{te}} (Y_i - \tilde{Y}_i)^2}{N_{te}}\}^{1/2}$) between Y and $\tilde{Y}$ were calculated to evaluate the prediction performance.

Spatial Outlier Detection. Nine different outlier detection approaches were applied to both simulation and real datasets. To compare the accuracies among them, we used the following common evaluation measures: detection rate (precision) and detection precision (recall). The precision was plotted against recall and the curves that are higher and farther to the right denote better performance.

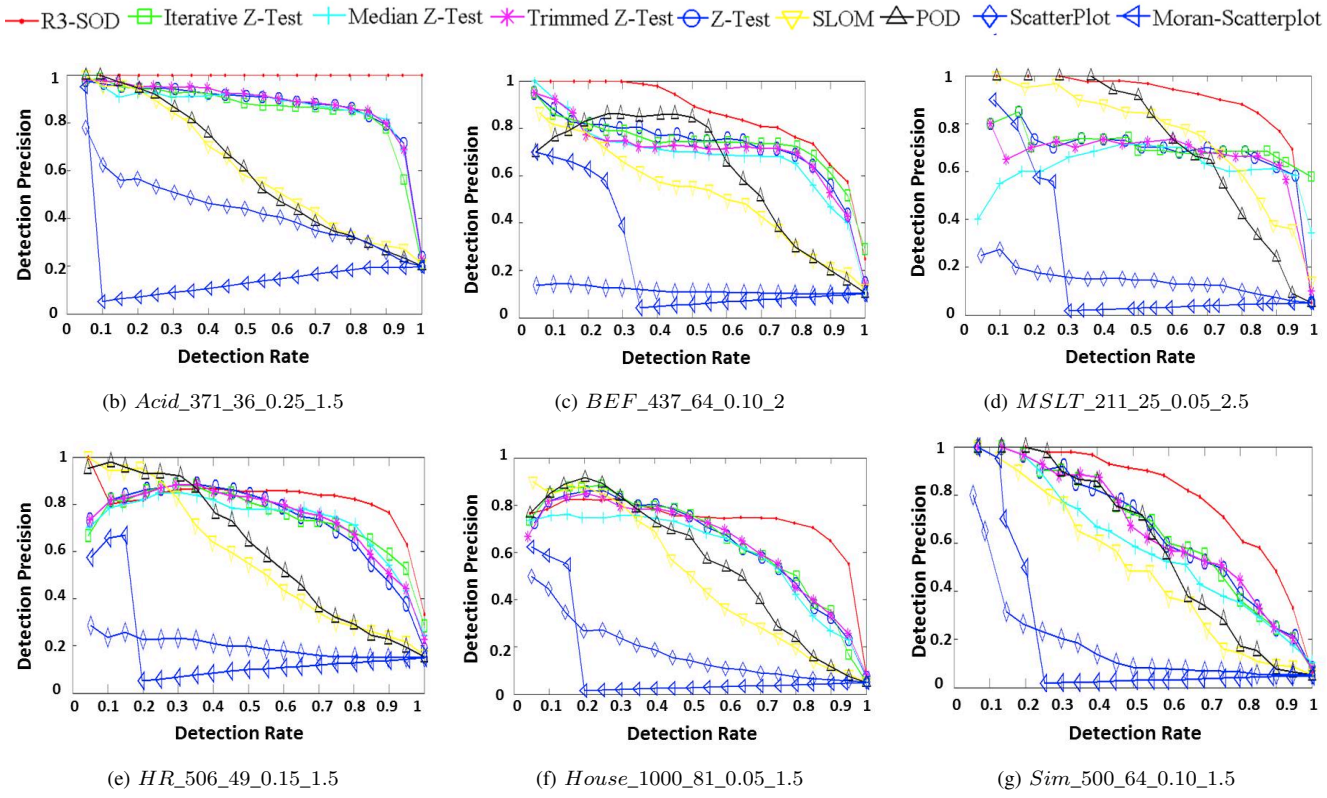
B. Experiment analysis and discussion

Robustness on Parameter Estimation. Table IV shows the parameter estimation results on four simulations with training data sizes of 300, 500, 700 and 1000. The data name depicts the parameter combination information. For example, "$Sim\_500\_100\_49\_0.05\_2.5$" indicates that it was generated by **simulation** data, and there are **500** training data, **100** testing data, **49** knots, and **5**% of the training data were contaminated as outliers by shifting the original Y to $(Y + \mathbf{2.5} * std(Y))$.

Comparing the estimated parameters with the true values, R^3-SKM was able to more accurately estimate most of the parameters. For "$Sim\_500\_100\_16\_0.05\_2.5$", only 5% the data are distorted with a relatively small shift rate(2.5), which means the contaminated data had a similar distribution to that of the original. Even so, SKM and Regress performed

Table IV: Comparison of parameter estimation results on simulations

Data	$Sim\_300\_30\_36\_0.25\_7$			$Sim\_500\_100\_49\_0.05\_2.5$			$Sim\_700\_200\_81\_0.05\_5$			$Sim\_1000\_400\_100\_0.05\_5$		
Para. / Approach	β	ϕ	σ^2	β	ϕ	σ^2	β	ϕ	σ^2	β	ϕ	σ^2
True Values	[0.50, 1.50]	25.00	2.00	[0.50, 1.50]	25.00	2.00	[0.50, 1.50]	25.00	2.00	[0.50, 1.50]	25.00	2.00
R^3-SKM	[0.49, 1.61]	24.73	1.11	[0.44, 1.54]	25.92	1.98	[0.48, 1.60]	26.41	1.80	[0.49, 1.46]	27.39	1.73
Regression	[0.44, 2.31]	–	–	[0.62, 1.84]	–	–	[0.39, 1.82]	–	–	[0.63, 1.31]	–	–
SKM	[0.06, 1.91]	6	1.85	[0.78, 1.23]	5.78	1.82	[0.29, 1.76]	5.66	1.77	[0.61, 1.38]	19077.44	48.33



(b) $Acid\_371\_36\_0.25\_1.5$ (c) $BEF\_437\_64\_0.10\_2$ (d) $MSLT\_211\_25\_0.05\_2.5$

(e) $HR\_506\_49\_0.15\_1.5$ (f) $House\_1000\_81\_0.05\_1.5$ (g) $Sim\_500\_64\_0.10\_1.5$

Figure 5: Comparison of SOD performances on simulation and real datasets

much worse than R^3-SKM on this data. As depicted by Table IV, when estimating β value, the estimation errors for R^3-SKM were 12% and 2.7% for β_1 and β_2, respectively. However, SKM had values of 56% and 18%, and Regression 24% and 22.6% for the same data. When estimating σ and ϕ, the estimation errors for R^3-SKM were 4% and 1%, while for SKM, they were 76.8% and 9%. This considerable difference in the estimation errors implies that SKM was highly influenced by the existence of outliers. There is no result for σ and ϕ for the Regression model, since this approach does not take the spatial dependency into consideration. And, as a consequence, it always incorrectly estimated β with higher errors, which absorbed the spatial variations into the spatial mean($X^T\beta$). Compared with SKM and Regression, R^3-SKM can be resilient to the influence of outliers, even in datasets in which more data were heavily contaminated, such as, "$Sim\_300\_30\_36\_0.25\_7$" in which 25% of data were skewed with a higher shift rate(7). Not surprisingly, some experimental results indicate that if the data is severely contaminated, it is more difficult to accurately estimate parameters. This is demonstrated by the lower performances of SKM and Regression in Sim_***_5** and Sim_***_7**. Still, R^3-SKM was able to achieve very impressive estimation results since the integration of the heavy tailed distribution helped alleviate the impact of outliers.

Robustness on Spatial Prediction. To better analyze prediction performances, we utilized Moran's I-statistic to capture the spatial dependency of Y observations. This made it possible to learn more about the degree of spatial autocorrelation for each dataset. The spatial dependency for Y in simulations was computed to be 0.70, which means that the simulations have a higher spatial dependency and this must be accurately captured during the estimation and prediction processes. The last column in Table III shows the calculated

spatial dependencies for real datasets. Most are lower values, which implies that the non-spatial attribute(X) contributes a great deal to the prediction the outcome variables.

Figure 4 compares the performances of different prediction models for simulated and real datasets. The calculated RMSEs and MAPEs demonstrate that R^3-SP outperforms both Regression and SKM. In particular, SKM did not perform as good as the Regression model in spite of the fact it takes into account the spatial auto-correlation in its spatial predictions. This is because the SKM model is considerably more complicated. It consists of a vector representing the spatial mean($x^T\beta$), the spatial random process(η) and measurement error ϵ. Regression is composed of $x^T\beta$ and ϵ. The hidden process η captures spatial association which is assumed to be a multivariate Gaussian process. When there are outliers in the dataset, SKM treats their outlying behaviors as natural spatial variations in the dataset, which therefore affects the computation of η, and further degrades the prediction quality. Meanwhile, the greater the outlying degrees of outliers, the worse its prediction performance: for the cases of HR_406_100_49_0.05_7, $MLST$_150_61_25_0.20_7 and Sim_300_30_36_0.25_7, where the shift rate are all 7. Interestingly, R^3-SP generated very similar prediction result to that for the Regression model for $Acid$_271_100_25_0.10_3.5. This is because Acid data has a very low spatial dependency, 0.007. So, spatial mean($x^T\beta$) dominates the prediction results. But for datasets with high spatial dependencies, like Sim_300_30_36_0.25_7 and $House$_1000_200_25_0.05_3, R^3-SP has preceding performance increases by benefiting from the integration of the heavy tailed distribution.

Accuracy of Spatial Outlier Detection. The outlier detection accuracies of different methods were compared based on different combinations of parameter settings. Figure 5 shows six representative results from simulated and real datasets. Clearly, R^3-SOD has impressive identification performances, achieving 10-15% improvement over Z, Median-Z, Iterative-Z and Trimmed-Z, 20-30% over POD and SLOM, 40-50% over Moran-Scatterplot, and 60-70% over Scatterplot. Z series of approaches identify outliers by normalizing the difference between a spatial object and the average of its spatial neighbors. However, this difference value is easily influenced by the presence of one or more outliers in its neighborhood, which leads to worse outlier detection qualities. This is especially true for higher numbers of outliers in the neighborhood, as demonstrated by "BEF_437_64_0.10_2" and "Sim_500_64_0.10_1.5". POD method constructs a graph based on k nearest neighbors, assigns the non-spatial attribute differences as edge weights, and then continuously cuts high weight edges to identify isolated points as outliers. Its performance degrades significantly with increasing outlier sizes. Such as in "$MSLT$_211_25_0.05_2.5", its performance was better than others since only 5% of the data were contaminated. The MoranScatterplot and Scatterplot approaches detect outliers by normalizing the attribute values against the average values for the corresponding neighborhood, which greatly neutralizes the significant differences caused by outliers and results in poor performances. It is worth mentioning that, if the outlying degrees of outliers are much higher(such as the shift rate is set to 4 and 5), all the SOD approaches can get good identification results. However, if the outlying behavior is less differentiated, they did not accurately capture spatial outliers at all except R^3-SOD. In contrast, R^3-SOD can be easily resilient to the outliers with different outlying degrees. When identifying outliers, it does not rely on neighborhood differences, which are susceptible to the neighborhood size and presence of outliers. Rather, it statistically analyzes the data model by integrating a heavy tailed distribution to minimize the effects of outliers. Its competing identification results are demonstrated by Figure 5.

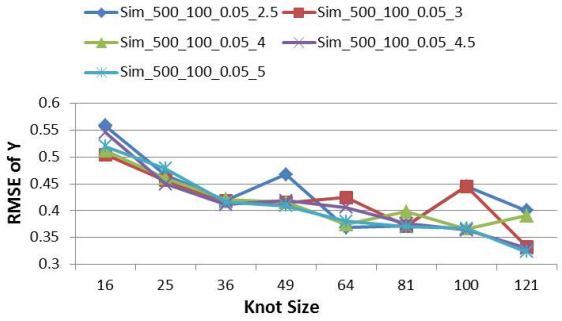


Figure 6: Prediction performances by varying knot sizes

Impact of Knot Sizes. We also evaluated the prediction performance by varying the knot sizes. Figure 6 shows various knot sizes from 16 to 121 (representing 3.2% and 24.2%, respectively, of the total number of observations). The curves show the effects of varying the number of knots on prediction accuracy. The simulations with different outlying degrees(2.5, 3, 4, 4.5, 5) have very similar affected trends for the different knot sizes. As shown in Figure 5, RMSEs decline as the knot size increases till they reach a stable state. In the simulations, which include 500 training and 100 testing points, the optimum prediction performance was achieved when knot size is equal to 64. That is, the knot size can be as high as 10%-15% of the total dataset. We also evaluated the optimal knot size in different sized datasets, and observed similar patterns. The optimal selection of the knot size enables not only more accurate spatial prediction, but also faster inferences.

Computational Cost. Finally, we examined the speed and associated scalability of SKM and R^3-SKM. Figure 7 displays the comparison of their runtime in datasets with varying numbers of training points. For all the simulations, the knot sizes were set to 10. Consequently, for R^3-SKM, when the data size is smaller than 1000, the time complexity

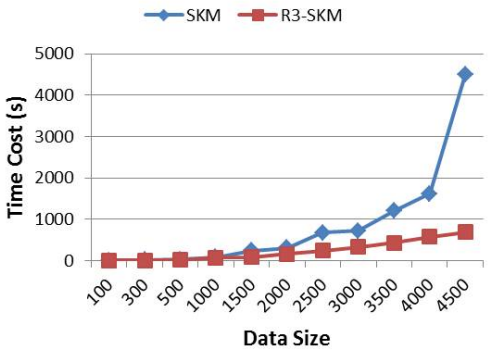


Figure 7: Total response time by varying data size

is dominated by the knot size, $O(m^3) = 1000$, rather than data size, $O(n)$. Above 1000, the time cost increases linearly. In comparison, the time cost of SKM was observed to increase in a nonlinear fashion for larger datasets. In summary, the reduced-rank techniques enables our algorithms to perform efficiently with a linear time complexity.

Result Discussion. R^3-SKM has been shown to be very robust for parameter estimation and spatial inference. It has superior performance over existing techniques in both real and simulation datasets. The experimental results verify three observations. First, if there is a good selection of knots that cover most of the domain interests, the predictive process cost will be significantly reduced to a linear order. Second, by being combined with numerical routines, Laplace approximation can provide much faster and more accurate parameter estimation. Third, integrating the heavy detailed distribution into the modeling process clearly minimizes the impact of outliers to a reasonable value, which provides a very good demonstration of the new method's robustness.

VII. Conclusion

This paper proposes a Robust and Reduced-Rank Spatial Kriging Model for large spatial datasets, abbreviated as R^3-SKM. This approach integrates a Bayesian hierarchical framework to support priors on model parameters. Meanwhile, the measurement error is modeled by a heavy tailed distribution, which enables it to be resilient to the influences of outliers and allow for fast spatial inferences. Furthermore, three algorithms are proposed to perform robust parameter estimation, spatial prediction and spatial outlier detection, respectively, in linear time. Their robustness and efficiency were demonstrated by extensive experimental evaluations. R^3-SKM provides critical functionality for stochastic processes on large spatial datasets.

References

[1] D. Higdon, "Space and space-time modeling using process convolutions," *Technique Report*, 2008.

[2] X. Lin, G. Wahba, D. Xiang, F. Gao, and R. K. M. B. Klein, "Smoothing spline anova models for large data sets with bernoulli observations and the randomized gacv," *Ann. Statist*, pp. 1570–1600, 2000.

[3] C. J. Paciorek, "Computational techniques for spatial logistic regression with large data sets," *Comput. Stat. Data Anal.*, vol. 51, no. 8, pp. 3631–3653, 2007.

[4] N. A. C. Cressie, *Statistics for Spatial Data*. Wiley-Interscience, 1993.

[5] S. Banerjee, A. E. Gelfand, A. O. Finley, and H. Sang, "Gaussian predictive process models for large spatial data sets," *Journal of the Royal Statistical Society Series B*, vol. 70, no. 4, pp. 825–848, 2008.

[6] R. A. Maronna, D. R. Martin, and V. J. Yohai, *Robust Statistics: Theory and Methods*. New York: John Wiley and Sons, 2006.

[7] J. Durbin and S. J. Koopman, "Monte carlo maximum likelihood estimation for non-gaussian state space models," *Biometrika*, vol. 84, pp. 669–684, 1997.

[8] P. Jylänki, J. Vanhatalo, and A. Vehtari, "Robust gaussian process regression with a student-t likelihood," *J. Mach. Learn. Res.*, vol. 12, pp. 3227–3257, 2011.

[9] M. Svensén and C. M. Bishop, "Robust bayesian mixture modelling," *Neurocomputing*, vol. 64, pp. 235–252, 2005.

[10] M. A. Gandhi and L. Mili, "Robust kalman filter based on a generalized maximum-likelihood-type estimator," *Trans. Sig. Proc.*, vol. 58, no. 5, pp. 2509–2520, May 2010.

[11] A. Y. Aravkin, B. M. Bell, J. V. Burke, and G. Pillonetto, "An 1 -laplace robust kalman smoother," *IEEE Trans. Automat. Contr.*, vol. 56, no. 12, pp. 2898–2911, 2011.

[12] S. Banerjee, A. E. Gelfand, A. O. Finley, and H. Sang. (2008) Gaussian predictive process models for large spatial data sets.

[13] H. Rue, S. Martino, and N. Chopin, "Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations," *Journal of the Royal Statistical Society*, vol. 71, no. 2, pp. 319–392, 2009.

[14] C. Varin, G. Host, and O. Skare, "Pairwise likelihood inference in spatial generalized linear mixed models," *Computational Statistics & Data Analysis*, vol. 49, no. 4, pp. 1173–1191, 2005.

[15] R. A. Dubin, "Spatial autocorrelation and neighborhood quality," *Regional Science and Urban Economics*, vol. 22, no. 3, pp. 433–452, 1992.

[16] A. O. Finley and S. Banerjee, "Hierarchical modeling for non-gaussian spatial data in r," 2009.

[17] spBayes, "spbayes: Univariate and multivariate spatial modeling," *http://cran.r-project.org/web/packages/spBayes/*, 2012.

[18] r. Björck, *Numerical methods for least squares problems*. SIAM, 1996.

[19] S. Shekhar, C.-T. Lu, and P. Zhang, "Detecting graph-based spatial outliers: algorithms and applications (a summary of results)," in *KDD*, 2001, pp. 371–376.

[20] C.-T. Lu, D. Chen, and Y. Kou, "Algorithms for spatial outlier detection," in *ICDM*, 2003, pp. 597–600.

[21] L. Anselin, "Local indicators of spatial association-lisa," *Geographical Analysis*, vol. 27, no. 2, pp. 93–115, 1995.

[22] R. Haining, *Spatial data analysis in the social and environmental sciences*. Cambridge University Press, 1990.

[23] B. Arunasalam, S. Chawla, P. Sun, and R. Munro, "Mining complex relationships in the sdss skyserver spatial database," in *COMPSAC Workshops*, 2004, pp. 142–145.

[24] Y. Kou, C.-T. Lu, and R. F. D. Santos, "Spatial outlier detection: A graph-based approach," in *ICTAI (1)*, 2007, pp. 281–288.