REVSTAT – Statistical Journal
Volume 18, Number 2, April 2020, 203–221

ROBUST ESTIMATION OF REDUCED RANK MODELS TO LARGE SPATIAL DATASETS

Authors: Casey M. Jelsema
– Department of Biostatistics, West Virginia University,
Morgantown, West Virginia, USA
jelsema.casey@gmail.com

Rajib Paul
– Department of Public Health Sciences, University of North Carolina – Charlotte,
Charlotte, North Carolina, USA
Rajib.Paul@uncc.edu

Joseph W. McKean
– Department of Statistics, Western Michigan University,
Kalamazoo, Michigan, USA
joseph.mckean@wmich.edu

Received: January 2019 Revised: September 2019 Accepted: September 2019

Abstract:

- For large datasets, spatial covariances are often modeled using basis functions and covariance of a reduced dimensional latent spatial process. For skewed data, likelihood based approaches with Gaussian assumption may not lead to faithful inference. Any L_2 norm based estimation is susceptible to long tails and outliers due to contamination. Our method is based on an empirical binned covariance matrix using the median absolute deviation and minimizes L_1 norm between empirical covariance and the model covariance. The consistency of the proposed estimate is established theoretically. The improvement is demonstrated using simulated data and cloud data obtained from NASA's Terra satellite.

Key-Words:

- *geostatistics; contaminated data; median absolute deviation; quantile regression; cloud.*

1. INTRODUCTION

Analysis of geostatistical data is known to be computationally intense or infeasible when the number of observed locations, n, is large. This is due to the size of the covariance matrix, $\mathbf{\Sigma}$ (which is $n \times n$) and the computational demand of inverting or factoring it. Cressie and Johannesson [4] introduced Fixed Rank Kriging (FRK) to address the computational hurdle by modeling the spatial covariance through a fixed number of deterministic basis functions and a latent reduced rank spatial process. To introduce the parameters, we consider an observed spatial process $Z(\mathbf{s})$ to be made up of a hidden spatial process $Y(\mathbf{s})$ along with a white noise process $\varepsilon(\mathbf{s})$ which could represent, for example, measurement errors. So we write

$$Z(\mathbf{s}) = Y(\mathbf{s}) + \varepsilon(\mathbf{s})\,. \tag{1.1}$$

Typically $Y(\mathbf{s})$ and $\varepsilon(\mathbf{s})$ are assumed to be independent Gaussian distributions, with $\varepsilon(\mathbf{s})$ having mean of zero. In this work however we develop methods that are robust to departure from this assumption. Then, for n observed locations, $Z(\mathbf{s}) \equiv \big\{Z(\mathbf{s}_1), ..., Z(\mathbf{s}_n)\big\}$ is an n-dimensional process with mean $E(Y(\mathbf{s})) = \boldsymbol{\mu}_Y$ and covariance matrix expressed as $\mathbf{\Sigma}_Z = \mathbf{\Sigma}_Y + \sigma^2\mathbf{I}_n$, where $\mathbf{\Sigma}_Y$ is the covariance matrix of $Y(\mathbf{s}) \equiv \big\{Y(\mathbf{s}_1), ..., Y(\mathbf{s}_n)\big\}$ and $\mathbf{I}_n$ is the identity matrix of rank n. We then model $Y(\mathbf{s})$ using a mixed effects model such as

$$Y(\mathbf{s}) = \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + \mathbf{S}(\mathbf{s})\boldsymbol{\eta} + \delta(\mathbf{s})\,. \tag{1.2}$$

In this model $\mathbf{X}(\mathbf{s})$ is a matrix of known covariates and $\boldsymbol{\beta}$ is the associated vector of regression coefficients; $\mathbf{S}(\mathbf{s})$ is a sparse $n \times r$ matrix of fixed, spatially varying basis functions which are centered at a set of r knot locations. Dimension reduction is achieved by selecting $r \ll n$. Various classes of basis functions may be used, including wavelets (Shi and Cressie [18] and Zhu *et al.* [22]) and bisquare (Cressie and Johannesson [4] and Paul *et al.* [16]) functions. The latent process $\boldsymbol{\eta}$ is a zero-mean r-dimensional Gaussian process defined over the knot locations, with covariance matrix $\mathbf{V}$. Finally $\delta(\mathbf{s})$, the process error, is an *iid* zero-mean Gaussian process with variance τ^2 which takes into account the variations unexplained by the large scale variations $\mathbf{X}(\mathbf{s})\boldsymbol{\beta}$ and spatial process $\mathbf{S}(\mathbf{s})\boldsymbol{\eta}$, and uncertainties arising from the dimension reduction. The process and measurement errors are usually assumed to be independent. When there is only one observation at each spatial location, τ^2 and σ^2 are non-identifiable, instead their sum $\nu^2 = \sigma^2 + \tau^2$, called the nugget variance, is estimated (though indirect means exist to estimate these separately, see Katzfuss and Cressie [11]). Going forward, we suppress the dependence on $\mathbf{s}$ when possible by stacking scalers into vectors, and vectors into matrices (e.g., $Y(\mathbf{s})$ is replaced with $\mathbf{Y}$ and $\mathbf{X}(\mathbf{s})$ is replaced with $\mathbf{X}$).

With this framework, the covariance matrix $\mathbf{\Sigma}_Z$ can be written as $\mathbf{\Sigma}_Z = \mathbf{SVS}' + \nu^2\mathbf{I}_n$. The objective is to estimate the model parameters: $\boldsymbol{\beta}, \mathbf{V}$ and ν^2. Once this has been done one may obtain the inverse of $\mathbf{\Sigma}_Z$ easily using the Sherman–Morrison–Woodbury matrix identity. This model offers a large degree of flexibility. The only restriction on $\mathbf{V}$ is the positive-definiteness, hence the resulting covariance matrix may be both anisotropic and nonstationary.

A variety of approaches have been used to model or estimate $\mathbf{V}$. In introducing FRK, Cressie and Johannesson [4] used a Method of Moments (MoM) estimation scheme, while Katzfuss and Cressie [11] developed an expectation-maximization (EM) algorithm. Much attention has also been given to Bayesian hierarchical modeling (see, for example, Banerjee *et al.* [1], Kang *et al.* [9] and Kang and Cressie [8]). To-date, little attention appears to have been given to robust estimation schemes. Zhu *et al.* [22] developed a method to reduce bias through improved basis function selection, but otherwise did not consider distributional assumptions. Paul *et al.* [16] developed a scale mixture model applicable to non-Gaussian datasets, but like many Bayesian methods it can be time-intensive to implement and run.

The basic FRK model we have described has been elaborated in various ways. For example, to obtain better representation of the spatial dependence some have used a tapering approach (Sang and Huang [17]) or multiple sets of knot locations with different resolutions (Cressie and Johannesson [4] and Kang *et al.* [10]). We demonstrate the latter approach in our data application in Section 5. Both the estimation and fitting stages in the existing MoM estimation use least-squares concepts, and therefore may suffer in the presence of skewed or contaminated data. In the present work we develop an alternative MoM estimator for the parameters of the RRSM. Our motivation in this is to provide an estimator that can model data containing outliers or exhibiting skewness, two features that are frequently encountered in geostatistical datasets, and which does not require significant computational resources.

MoM estimation of the model parameters is divided into two stages: an estimation stage and a fitting stage. In the estimation stage, the entire spatial domain is divided into M bins such that $r < M \ll n$, and $\boldsymbol{\Sigma}_M$ is defined to be the covariance matrix over the bins. The bins are defined subjectively, though Cressie and Johannesson [4] and Katzfuss and Cressie [11] provide some recommendations. Then an empirical estimate $\hat{\boldsymbol{\Sigma}}_M$ is constructed using the *detail residuals*, $\mathbf{D} = \mathbf{Z} - \mathbf{X}\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the ordinary least squares estimate of $\boldsymbol{\beta}$. Cressie and Johannesson [4] defined $\hat{\boldsymbol{\Sigma}}_M$ in the following manner: The m^{th} diagonal elements $\hat{\boldsymbol{\Sigma}}_M(m,m) = \text{avg}(\mathbf{D}_m^2)$ and the (m,m') off-diagonal element $\hat{\boldsymbol{\Sigma}}_M(m,m') = \text{avg}(\mathbf{D}_m) \times \text{avg}(\mathbf{D}_{m'})$. In these expressions, $\mathbf{D}_m$ is the vector of detail residuals in bin m, and $\text{avg}(\cdot)$ denotes the average.

Similarly $\mathbf{S}$ is binned into an $M \times r$ matrix by taking the column averages of the rows of $\mathbf{S}$ associated with the observed locations falling into each of the M bins. Denoting this as $\overline{\mathbf{S}}$, one may then write

$$\boldsymbol{\Sigma}_M = \overline{\mathbf{S}}\mathbf{V}\overline{\mathbf{S}}' + \nu^2 \mathbf{I}_M \,. \tag{1.3}$$

After estimation, the fitting stage obtains $\hat{\mathbf{V}}$ and $\hat{\nu}^2$ by minimizing the Frobenius norm between $\boldsymbol{\Sigma}_M$ and $\hat{\boldsymbol{\Sigma}}_M$, using the QR-decomposition on $\overline{\mathbf{S}}$. This is a two-step process resulting in the following estimates:

$$\begin{aligned} \hat{\nu}^2 &= (\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'\big(\hat{\boldsymbol{\Sigma}}_M - \mathbf{Q}\mathbf{Q}'\hat{\boldsymbol{\Sigma}}_M\mathbf{Q}\mathbf{Q}'\big)\,, \\ \hat{\mathbf{V}} &= \mathbf{R}^{-1}\mathbf{Q}'\big(\hat{\boldsymbol{\Sigma}}_M - \hat{\nu}^2\mathbf{I}_M\big)\mathbf{Q}\mathbf{R}'^{-1}\,, \end{aligned}$$

where $\mathbf{F} = \mathbf{I}_M - \mathbf{Q}\mathbf{Q}'$. If $\hat{\boldsymbol{\Sigma}}_M$ is not positive-definite, the eigenvalues must be lifted to ensure that $\hat{\mathbf{V}}$ is positive-definite (see Kang *et al.* [9]). For further details on Fixed Rank Kriging, see Katzfuss and Cressie [11].

We redesign both the estimation and fitting stages for the MoM estimation using the Median Absolute Deviation and quantile regression (Section 2). Our work is novel in that we return to basic principles to redesign the estimation and fitting stages with a mind for resisting contaminated data. The consistency of our proposed estimate is shown (Section 3), though the technical details are given in the Appendix. We describe and conduct a simulation study (Section 4) to investigate the performance of our proposed method. Finally, we provide a data example (Section 5) using a large remote sensing dataset and some concluding remarks (Section 6).

2. ROBUST ESTIMATION AND FITTING

In this section we describe robust alternatives to both the estimation stage and fitting stage of MoM estimation for the FRK model. First we define $\hat{\mathbf{\Sigma}}_M^{(\text{rob})}$ as an estimate empirical binned covariance matrix which is robust to contamination. Then we describe a robust strategy to fit the model parameters, which we call the *robust fit*. We denote the previous-described methods from Cressie and Johannesson [4] as $\hat{\mathbf{\Sigma}}_M^{(\text{CJ})}$ and the Frobenius fit.

2.1. Estimation stage

The diagonal elements of $\mathbf{\Sigma}_M$ represent the variance within a bin. We estimate this quantity using the median absolute deviation, $\text{MAD}(X) = \text{med}\big(|X - \text{med}(X)|\big)$. A constant scale factor is applied to the MAD which causes it to be a consistent estimate for the standard deviation (see Hettmansperger and McKean [7], Eqn. 3.9.27). In the present work, we use the usual MAD which is consistent for σ when the errors are normally distributed. Hence, the diagonal elements of our proposed estimate are given by

$$\hat{\mathbf{\Sigma}}_M^{(\text{rob})}(m,m) = \text{MAD}^2(\mathbf{D}_m)\,, \qquad m = 1, ..., M\,. \tag{2.1}$$

Estimating the covariance between two bins is more challenging. First, recall that $\text{cov}(A,B) = \frac{1}{4}\big[V(A+B) - V(A-B)\big]$. Estimating a covariance using this identity requires finding $\mathbf{D}_m \pm \mathbf{D}_{m'}$, however, these quantities are not well-defined. For example, two bins may not even have the same number of observations, much less any natural correspondence between observations. We therefore use the pairwise sums and pairwise differences, denoted using $\oplus$ and $\ominus$ respectively, to approximate $\mathbf{D}_m \pm \mathbf{D}_{m'}$. We again use the square of the MAD to estimate the variance, so the off-diagonal elements of our estimate are given by:

$$\hat{\mathbf{\Sigma}}_M^{(\text{rob})}(m,m') \;=\; \frac{1}{4}\Big[\text{MAD}^2(\mathbf{D}_m \oplus \mathbf{D}_{m'}) - \text{MAD}^2(\mathbf{D}_m \ominus \mathbf{D}_{m'})\Big]. \tag{2.2}$$

2.2. Fitting stage

Given an empirical covariance matrix $\hat{\mathbf{\Sigma}}_M$, we fit $\mathbf{V}$ by minimizing some norm between $\hat{\mathbf{\Sigma}}_M$ and $\mathbf{\Sigma}_M$. To develop the robust fitting stage, we start from equation (1.3),

$$\begin{aligned} \hat{\mathbf{\Sigma}}_M &= \overline{\mathbf{S}}\mathbf{V}\overline{\mathbf{S}}' + \hat{\nu}^2\mathbf{I}_M\,, \\ \big(\hat{\mathbf{\Sigma}}_M - \hat{\nu}^2\mathbf{I}_M\big)\,\overline{\mathbf{S}}\big(\overline{\mathbf{S}}'\overline{\mathbf{S}}\big)^{-1} &= \overline{\mathbf{S}}\mathbf{V}. \end{aligned} \tag{2.3}$$

Then we may see equation (2.3) as a multivariate regression problem with $\overline{\mathbf{S}}$ as the design matrix and $\mathbf{V}$ as the matrix of regression coefficients. Any method of robust regression may then be implemented to obtain an estimate of $\mathbf{V}$. For this work, we use the popular least absolute deviations, L_1, estimator; see Koenker and Bassett [13] and Section 3.8 of Hettmansperger and McKean [7]. In comparison to least squares (LS), the least absolute deviation fit is obtained by replacing the squared Euclidean norm with the L_1 norm. Hence, the geometry and interpretation of the L_1 fit is quite similar to LS fit, but unlike the LS estimate, the L_1 estimate is robust. As discussed in Section 3.8 of Hettmansperger and McKean [7], the fit is also efficient. It attains efficiency 0.64 relative to LS for normal errors but is generally more efficient than LS for error distributions with tails heavier than the normal.

Each column of $\left(\hat{\mathbf{\Sigma}}_M - \hat{\nu}^2\mathbf{I}_M\right)\overline{\mathbf{S}}\left(\overline{\mathbf{S}}'\overline{\mathbf{S}}\right)^{-1}$ is used as the response in a separate estimation. There are therefore r estimates to obtain, each of which corresponds to a column of $\mathbf{V}$. As the final estimate $\mathbf{V}$ may not be numerically symmetric, we symmetrize $\hat{\mathbf{V}}$ by taking $\hat{\mathbf{V}} = 0.5\left(\hat{\mathbf{V}} + \hat{\mathbf{V}}'\right)$. We used the `quantreg` `R` package (Koenker [12]) for the computation of the L_1 fit.

Estimation of $\mathbf{V}$ requires an estimate of ν^2. By substituting the left side of (2.3) for $\overline{\mathbf{S}}\mathbf{V}$ in (1.3) we obtain:

$$\begin{aligned}\hat{\mathbf{\Sigma}}_M &= \left(\hat{\mathbf{\Sigma}}_M - \nu^2\mathbf{I}_M\right)\overline{\mathbf{S}}\left(\overline{\mathbf{S}}'\overline{\mathbf{S}}\right)^{-1}\overline{\mathbf{S}}' + \nu^2\mathbf{I}_M\,,\\ \hat{\mathbf{\Sigma}}_M\Big(\mathbf{I}_M - \overline{\mathbf{S}}\left(\overline{\mathbf{S}}'\overline{\mathbf{S}}\right)^{-1}\overline{\mathbf{S}}'\Big) &= \nu^2\Big(\mathbf{I}_M - \overline{\mathbf{S}}\left(\overline{\mathbf{S}}'\overline{\mathbf{S}}\right)^{-1}\overline{\mathbf{S}}'\Big)\,. \end{aligned} \tag{2.4}$$

We then stack the columns of $\hat{\mathbf{\Sigma}}_M\left(\mathbf{I}_M - \overline{\mathbf{S}}\left(\overline{\mathbf{S}}'\overline{\mathbf{S}}\right)^{-1}\overline{\mathbf{S}}'\right)$ and the columns of $\left(\mathbf{I}_M - \overline{\mathbf{S}}\left(\overline{\mathbf{S}}'\overline{\mathbf{S}}\right)^{-1}\overline{\mathbf{S}}'\right)$. Doing this, we again cast the problem as a zero-intercept robust regression, where ν^2 is the slope. This estimate is substituted into equation (2.3) to obtain an estimate of $\mathbf{V}$.

The estimate of $\mathbf{V}$ may not be positive-definite, so we may need to lift the eigenvalues (similar to Cressie and Johannesson [4]), while preserving the total variability. In our work, we compute the sum of the eigenvalues, Δ, and proportionally redistribute this sum across the eigenvalues after shifting all eigenvalues to be non-negative.

3. ASYMPTOTIC PROPERTIES

Here we discuss some of the infill asymptotic properties of our proposed estimator, $\hat{\mathbf{\Sigma}}_M^{(\mathrm{rob})}$. Infill asymptotics is a common method of considering asymptotics related to geostatistical methodology in which the domain, $\mathcal{D}$, remains fixed but the density of observed locations is increased.

Recall that we obtain $\hat{\mathbf{V}}$ by minimizing some norm $\|\cdot\|$:

$$\hat{\mathbf{V}} = \operatorname{argmin}\left\|\hat{\mathbf{\Sigma}}_M - \mathbf{\Sigma}_M\right\|.$$

Hence, once $\hat{\mathbf{\Sigma}}_M$ is known, $\hat{\mathbf{V}}$ is fully determined by the fitting method. Therefore, a desirable property of the empirical binned covariance matrix $\hat{\mathbf{\Sigma}}_M^{(\mathrm{rob})}$ is that it be consistent for $\mathbf{\Sigma}_M$, which we establish in this section.

There are two sets of assumptions that we need to make. From expressions (2.1) and (2.2), $\hat{\mathbf{\Sigma}}_M^{(\mathrm{rob})}$ is a function of MADs applied to the detail residuals. For each bin m, these residuals are obtained from ordinary least-squares regression, our proof requires that $\sqrt{n}\,\big(\hat{\beta}-\beta\big) = O(1)$ for each bin. For this, we assume the conditions in the paper by Lahiri *et al.* [14] for each bin.

Our process for bin j (slightly abusing the notation to avoid double subscript), is $\big\{e_1, e_2, ..., e_{n_j}\big\}$ which we denote by $\{\mathbf{e}_j\}$. On this process we assume that

1. $\{\mathbf{e}_j\}$ is stationary.

2. $\{\mathbf{e}_j\}$ satisfies the strong mixing coefficients assumption given as follows. For $i \neq k$, let A_i and B_k be in the σ-fields generated by e_i and e_k. Then

$$\Big|P[A_i \cap B_k] - P[A_i]\,P[B_k]\Big| = O\big(\rho^{|i-k|}\big)\,, \tag{3.1}$$

where $0 \leq \rho < 1$.

Note that Assumption 2 implies that the spatial correlation between two locations exhibits exponential decay. This is a common feature in spatial modeling (e.g. the Matérn class of covariance models), and as such is not an unreasonable assumption.

For our proof, let $\mathbf{D}_m$ denote the random detail residual process within the m^{th} bin, and let $\mathbf{D}_m = \big\{\tilde{R}_{m_1}, ..., \tilde{R}_{m_k}\big\}$ be the k observed detail residuals from that bin. We assume that $\mathbf{D}_m$ and, as will be seen, $|\mathbf{D}_m|$, exhibit strong mixing as described in conditions 1 and 2.

We now state the consistency result in theorem form. The proof is given in the Appendix.

Theorem 3.1. *Under the above conditions,* $\hat{\mathbf{\Sigma}}_M^{(\mathrm{rob})}$ *is a consistent estimator of* $\mathbf{\Sigma}_M$*.*

Throughout we treat the number of bins, M, as fixed, and do not consider limits over that quantity. This is analogous to the work of Bliznyuk *et al.* [2]. In another context on binned estimation, they considered m (the number of bins) as a radius to determine "adjacency" of locations, where m does not depend on n, (the number of observations) and did not limit over m. The only restriction on M is that it should be large enough to ensure that the assumption of stationary within bins holds for practical implementation.

4. SIMULATION STUDY

To compare our proposed methods with the existing methods using simulated data, we generate a spatial process Z according to the model:

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{S}\boldsymbol{\eta} + \boldsymbol{\epsilon}\,.$$

First we select n locations uniformly over a 100×100 domain, and $r_o = 1225$ knot locations on a 35×35 grid. These knot locations are used to simulate the data but not to fit the models (because reduced rank spatial models are designed as approximations of a more complex spatial process). Then we define $\mathbf{X}$ as an $n\times3$ matrix where the columns correspond respectively to an intercept, the x-coordinate, and the y-coordinate.

To define $\mathbf{V}$ we first compute the pairwise distances between the knot locations, and generate a Matérn covariance matrix using these distances with sill and range parameters each set to 1, and smoothness set to 0.5. We use `cov.sp` in the `R` package `SpatialTools` (French [6]) to generate this matrix. We then obtain $\mathbf{V}$ as an observation from the inverse Wishart distribution using the Matérn covariance as a scale matrix and $2(r+1)$ degrees of freedom. In this way the covariance matrix used to simulate the data is not constrained to be either stationary or isotropic.

We construct $\mathbf{S}$ using the bisquare basis functions defined as

$$S_{i,j} = \begin{cases} \left(1 - \left(\|s_i - u_j\|/r_u\right)^2\right)^2 & \text{for } \|s_i - u_j\| \le r_u\,, \\ 0 & \text{otherwise}\,, \end{cases}$$

where r_u is 1.5 times the minimum distance between knots and $\|\cdot\|$ denotes the measure of distance appropriate to the data (e.g., in our simulations, we used Euclidean distance).

We used two methods to simulate the data, a Contaminated Normal distribution and an Exponential distribution. These simulate the presence of outliers or of skewness, respectively, in the resulting dataset. For either simulation method, we compare the model fits by splitting the simulated data into a training set and a held-out test set. The hold-out set was set as all of the locations in the square bounded by the points $(40, 40)$ and $(60, 60)$, which corresponds to approximately 4% of the observations. We use the estimated parameters to predict at the held-out locations and compute diagnostics to assess both the accuracy and uncertainty of the prediction, including the mean square error (MSE), mean square prediction error (MSPE), and the continuous ranked probability score (CRPS, Wilks [21]), a measure which incorporates both the prediction accuracy and the prediction uncertainty. Lower values are preferable for all of these measures.

4.1. Simulation 1: contaminated normal

For simulating datasets we first generate a r_o-dimensional process $\boldsymbol{\eta}$ from a zero-mean multivariate normal with covariance $\mathbf{V}$. To induce outliers, the measurement error process $\boldsymbol{\varepsilon}$ is generated from a contaminated normal distribution. We first draw a random sample from $\mathcal{N}(0, \nu^2)$, and then replace αn of the values with random draws from $\mathcal{N}(0, \nu_c^2)$. Finally, we obtain the simulated data by $\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{S}\boldsymbol{\eta} + \boldsymbol{\varepsilon}$. For each simulated dataset, estimate model parameters using both the method of Cressie and Johannesson [4] and the proposed robust method.

We considered three sample sizes, $n \in \{10000, 15000, 20000\}$ and five levels for the number of knots locations to fit the model, $r \in \{64, 100, 144, 196, 256\}$, intentionally chosen to much less than r_o, so that the "true" spatial process was more granular than the model. For the contamination level of $\boldsymbol{\varepsilon}$ we consider $\alpha \in \{0.00, 0.05, 0.10, 0.15, 0.20\}$. For the simulations shown, the values of $\boldsymbol{\beta} = (1, 0.01, 0.05)'$, $\nu^2 = 1$, and $\nu_c^2 = 100$ were held constant. These choices are not sensitive to our estimation technique except insofar as a larger or smaller ν_c^2 would correspond to a larger or smaller effect from the contamination. For each combination of these parameters, we generated 50 replications of data. Hence, there were 75 settings of parameter levels, and 3750 replications in total.

4.2. Simulation 2: exponential

As we have noted throughout, skewness can also be problematic for least-squares type estimators, and skewed data are not uncommon in geostatistics. Hence, we designed a second simulation in which we generate $\boldsymbol{\varepsilon}$ from an Exponential distribution rather than from a contaminated Normal distribution. We use the same design as Simulation 1, but instead of α, we consider the rate parameter of the Exponential distribution $\lambda \in \{0.10, 0.25, 0.50, 1.00\}$. Hence, for this simulation there were 60 settings and 3000 replications in total.

4.3. Simulation results

The simulations suggest that the robust method is generally preferable to the CJ method. For brevity we present the results for the CRPS, but results for the MSE and MSPE were similar. We use two main values to compare the results: The median CRPS across the 50 replications, and the CRPS of the CJ method relative to that of the robust method (we refer to this as the CRPS ratio).

Results of Simulation 1 are shown in Figure 1, which plots the median CRPS over the 50 replications for each of the settings. In 67 of the 75 settings, the robust method produced a smaller median CRPS than the CJ method.

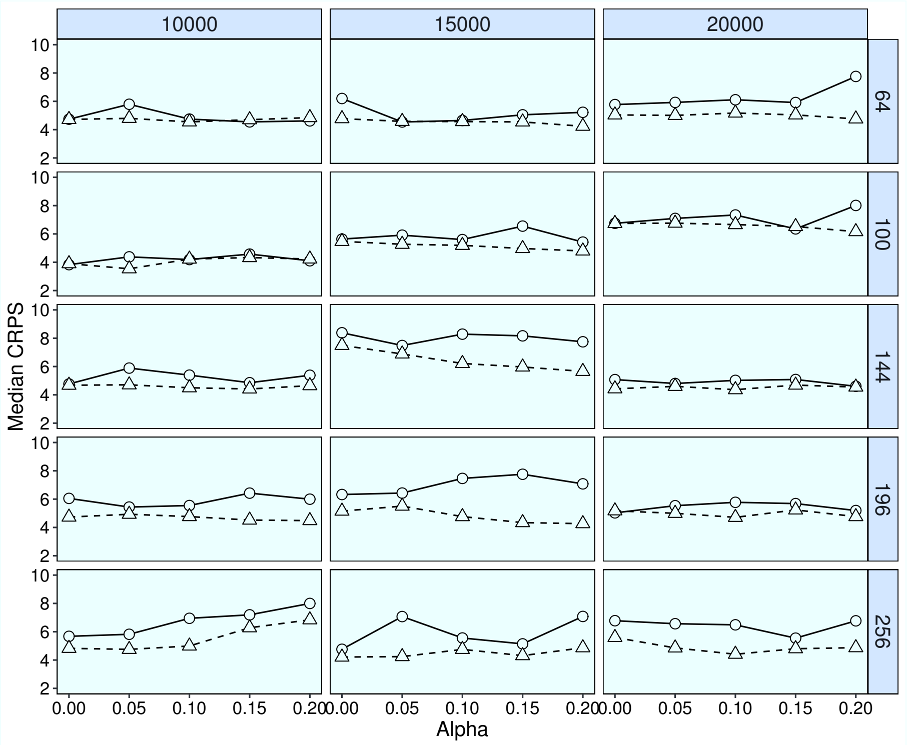

Figure 1: Results for Simulation 1. Plotted points are median CRPS of the CJ method (circles) and the robust method (triangles) over the 50 replications.

In addition, the robust method produced a smaller CRPS (i.e. CRPS ratio greater than 1) in 68.8% of the replications, and the median of the CRPS ratio showed a 9% larger CRPS for the CJ method. When considering the CRPS ratio for each setting, the worst-performing setting for the robust method had a median CPRS ratio of 0.975 (near equivalence), while half of the settings had a median CRPS ratio showing an improvement of 10% or more.

The results for Simulation 2 were similar to those of Simulation 1, and are shown in Figure 2. In 55 of the 75 settings, the robust method produced a smaller median CRPS than the CJ method.

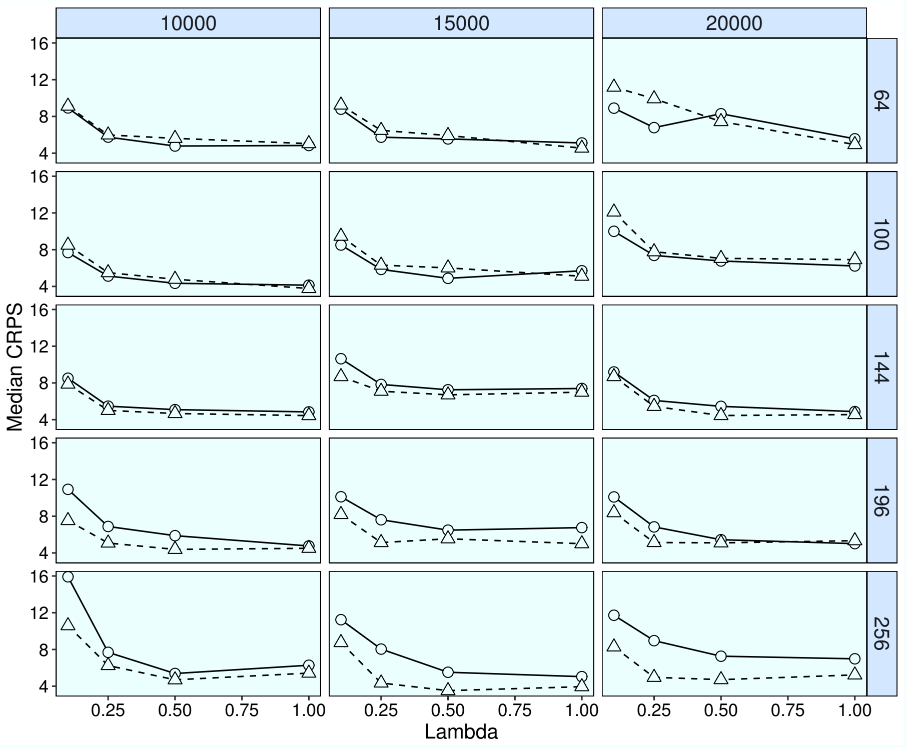

Figure 2: Results for Simulation 2. Plotted points are median CRPS of the CJ method (circles) and the robust method (triangles) over the 50 replications.

In addition, the robust method produced a smaller CRPS (i.e. CRPS ratio greater than 1) in 65.3% of the replications, and the median of the CRPS ratio showed an 8% larger CRPS for the CJ method. When considering the CRPS ratio for each setting, the worst-performing setting for the robust method had a median CPRS ratio of 0.957, which again shows minimal advantage for the CJ method, while half of the settings had a median CRPS ratio showing an improvement of at least 7%.

To provide an overall summary of our results, our findings suggest that the proposed robust method tends to be advantageous compared to the CJ method. While we acknowledge this is not uniformly the case, we note that in approximately two-thirds of cases, the proposed method resulted in smaller CRPS. It is unfortunately difficult to discern much in the way of a pattern across the simulation settings, to determine whether the robust or CJ method might be preferable in a specific setting. The main apparent pattern from these simulations is that the more knots, the better the robust method tended to perform against the CJ method.

This could potentially be a consequence of each bin from the estimation of $\mathbf{\Sigma}_M$ having fewer observations compared to a setting with the same sample size but smaller number of knots, in which case outliers would have an increased effect.

Since the number of knots is chosen by the modeler, one might be tempted to select a smaller value of r, so that any effect from the choice of method is minimized. However, fewer knots corresponds to a more coarse representation of the spatial variation, hence the general recommendation (e.g. Finley *et al.* [5]) is to use as many as possible (within any computational limits). Hence, the natural choice guiding the selection of r will also tend to produce situations in which the robust method appears to perform better.

5. APPLICATION TO NASA DATA

We use remote sensing data on daily cloud liquid water path (CWP), obtained through NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) on the Terra satellite on April 22, 2012. Note that this date is an arbitrary choice, our interest here is to demonstrate our method outside of a fabricated example. Because the dataset is large ($n = 48552$), a reduced rank model is a reasonable choice for inference. The CWP data are right-skewed, so we restrict out focus to the log-scale.

5.1. Original Data Analysis

The observed data are plotted in Figure 3. Due to a north-south trend (tending to smaller values closer to the equator), we model the large-scale variation using Legendre polynomials similar to Stein [19], though using only the latitude. Specifically, let L denote the degrees latitudes and define $\ell = \pi(L/180)$.

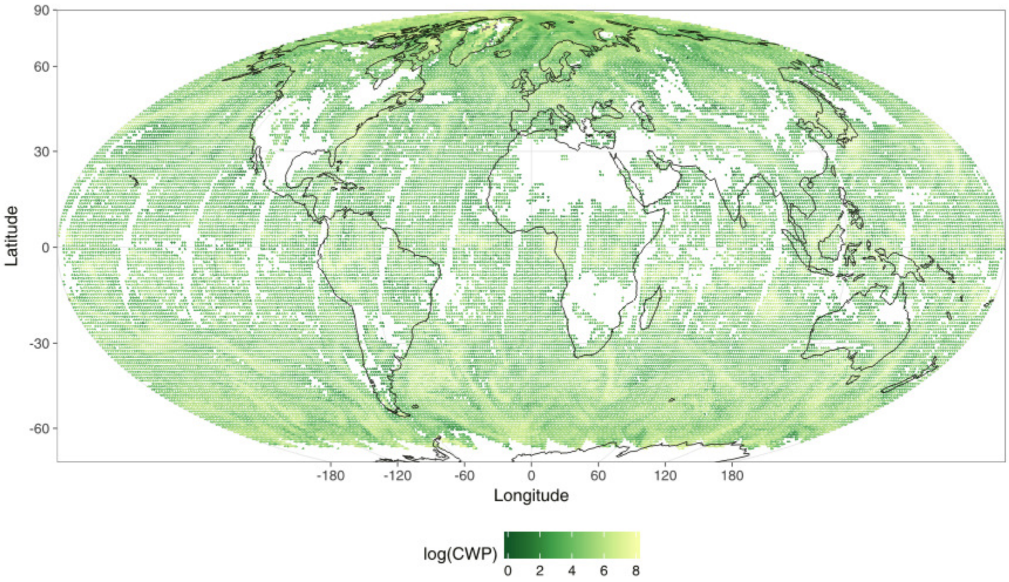

Figure 3: Plot of observed Cloud Water Path over the spatial domain.

We compute Legendre polynomials $P_p^q(\sin(\ell))$ of degree $p = 80$ and order $q = 0, 1, ..., p$. This results in a design matrix consisting of 81 regressors of spherical harmonics. Stein [19] also included a cosine of the longitude. Since we observed primarily a trend over the latitudes, we do not include the cosine term on longitude. Since our focus is on the small-scale (spatial) variation rather than the large-scale variation, the main concern for us is that this model enables stationarity of the spatial process to be reasonable; visual inspection (figure not shown) of the predictions for each latitude show this to be the case.

For the MoM estimation described in the preceding sections we first compute the detailed residuals. The normal quantile-quantile plot of the detailed residuals in Figure 4 shows a heavy lower tail, which motivates the use of the proposed robust techniques. Initially we model the data as observed. Afterwards, we also induce outliers into the data and reanalyze the data.

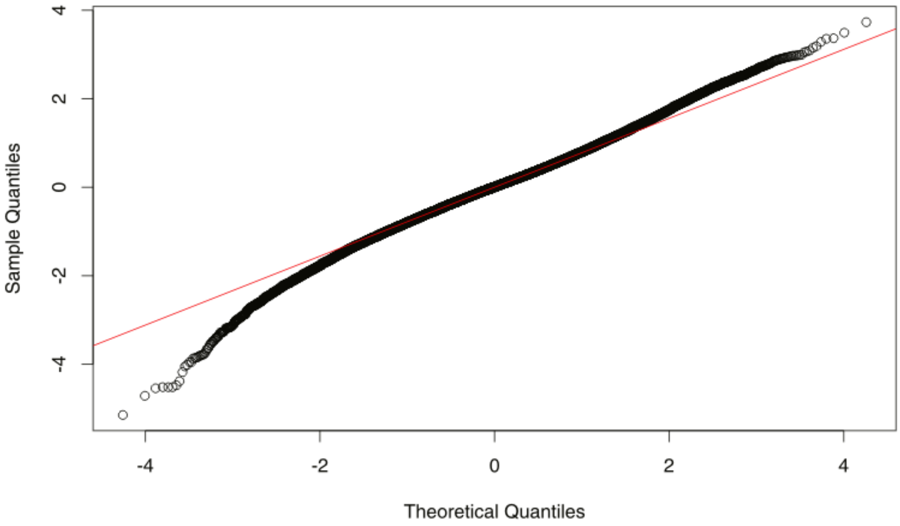

Figure 4: Normal quantile-quantile plot of the detailed residuals.

As recommended by Cressie and Johannesson [4], we use a multi-resolution model for CWP (see Nychka *et al.* [15]), to capture multiple scales of variation. We choose $r_1 = 38$ knot locations for the first resolution, and $r_2 = 97$ knot locations for the second resolution. Therefore the estimate of $\mathbf{V}$ is a 135×135 matrix. A map of these knot locations is given in Figure 5.

To construct the $\mathbf{S}$ matrix, we use the modified bisquare function, defined as:

$$\mathbf{S}_{i,j(l)} = \begin{cases} \Big(1 - 0.25\, d^2\big(\mathbf{s}_i, \mathbf{u}_{j(l)}\big)\Big) & \text{for } d\big(\mathbf{s}_i, \mathbf{u}_{j(l)}\big) \le 2\,, \\ 0 & \text{otherwise}\,, \end{cases}$$

where $\mathbf{u}_{j(l)}$ is the j^{th} knot location of the l^{th} resolution, $\mathbf{s}_i$ are the observed locations. The distance is given by:

$$d\big(\mathbf{s}_i, \mathbf{u}_{j(l)}\big) = \sqrt{d^2_{\text{long}}\big(\mathbf{s}_i, \mathbf{u}_{j(l)}\big)/r^2_{\text{long}(l)} + d^2_{\text{lat}}\big(\mathbf{s}_i, \mathbf{u}_{j(l)}\big)/r^2_{\text{lat}(l)}}\;,$$

where $d_{\text{long}}\big(\mathbf{s}_i, \mathbf{u}_{j(l)}\big)$ and $d_{\text{lat}}\big(\mathbf{s}_i, \mathbf{u}_{j(l)}\big)$ denote the longitude (east-west) and latitude (north-south) distances, respectively, between the location $\mathbf{s}$ and the knot location $\mathbf{u}_{j(l)}$. The values

$r_{\text{long}(l)}$ and $r_{\text{lat}(l)}$ control the maximum distance between an observation and a knot such that there is non-zero weight associated between the two. We set these to be the minimum east-west distance and minimum north-south distance between two knot locations of the same resolution.

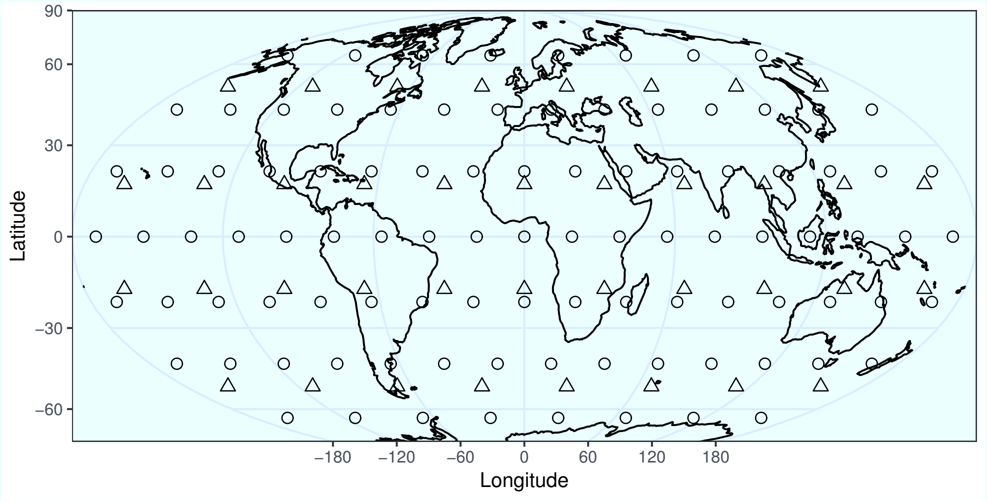

Figure 5: Plot of the knot locations of the basis functions over the spatial domain. Triangles represent the 38 knot locations of the first resolution, and circles represent the 97 knot locations of the second resolution.

Figures of the predictions or prediction uncertainties are not particularly informative, as our focus is on comparing the robust method to the CJ method. The CJ method yielded larger RMSPEs by approximately 20%, and the CRPS tended to be larger as well. A plot of the CRPS ratio for each location is shown in Figure 6. On average, the CRPS ratio is 1.04, indicating better performance for the robust method.

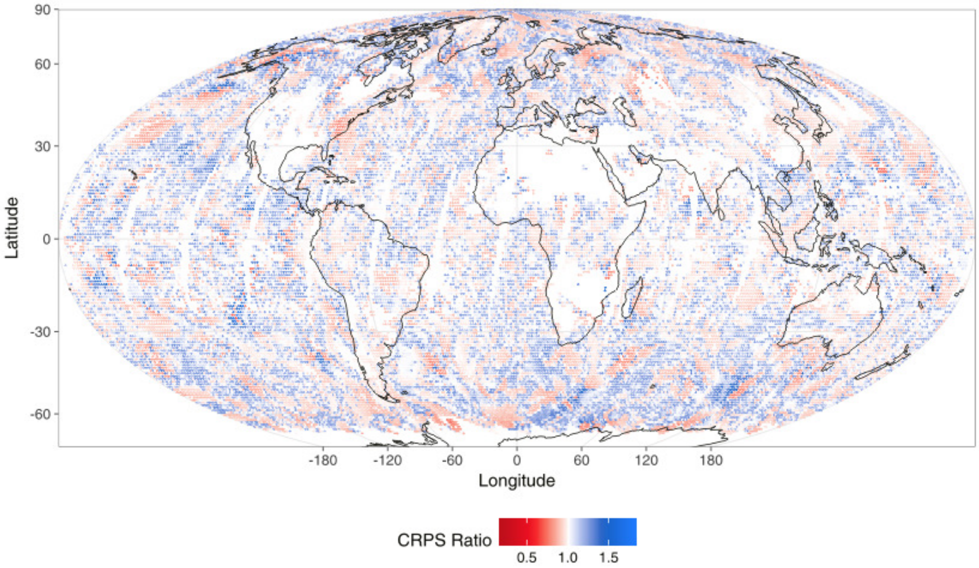

Figure 6: Plot of the CRPS of predictions using the CJ method relative to those using the robust method. Larger values indicate the CJ method produced a larger CRPS at that location.

5.2. Analysis after inducing outliers

In addition to this analysis, we artificially contaminated the log CWP data by replacing the 2% of observed values $Z_i(\mathbf{s})$ with $1.5\, Z_i(\mathbf{s})$. Inspection of the normal quantile-quantile plot showed a heavy upper tail which also contained many outliers. The results followed the same pattern as those described above. The RMSPE were again uniformly larger for the CJ method, now averaging 78% larger, while the CRPS were, on average, 11% larger.

6. CONCLUSIONS AND DISCUSSION

The Method of Moments is a flexible and powerful tool for estimating the parameters of a FRK model. Bayesian methods are more accurate than kriging (Kang and Cressie [8]), but they are also more time-consuming, and often come with some distributional assumptions. Kriging is typically a faster process, and kriging estimates are BLUP even in the face of non-normality, so kriging presents benefits of its own. However the typical parameter estimates using EM algorithm or MoM are susceptible to contaminated data. In this work we have provided robust alternatives to both stages of the MoM estimation.

Our results indicate that the proposed estimate and fitting scheme successfully capture the spatial covariance. In both our simulations and in our application to real data, the robust method tended to provide an advantage over the CJ method. At times the advantage was small, but in some cases the robust method showed substantial improvement, even when the data were neither contaminated or skewed.

Besides the L_1-fit, other robust fits can be used. For example, the Wilcoxon fit is a robust fit that minimizes the sum of the absolute differences of the residuals (see Hettmansperger and McKean [7], Section 3.8). The Wilcoxon fit is generally more efficient than the L_1-fit and it generalizes to fits for skewed-error distributions. We are currently investigating other robust norms which result in fits with higher efficiency than that of the L_1 fit for normal errors.

Again we emphasize that the kriging equations have been derived by minimizing the mean square prediction error. These predictions are then simply functions of $\mathbf{V}$ and ν. In our work, we have provided robust methods of estimating these same parameters. Yet when using robust techniques, it may be desirable to derive predictions and measures of precision using a different loss function than the squared error loss, or such that the predictions are robust in addition to the parameter estimates (Cressie and Hawkins [3]). Our robust estimates perform well in spite of this.

A. APPENDIX – Proof of Theorem 3.1

The proof utilizes the consistency of a fit $\hat{\beta}$ such that $\sqrt{n}\,(\hat{\beta}-\beta)=O(1)$; the assumptions as discussed in Section 3, including $n_j\to\infty$, for $j=1,...M$; and the theory for the sign processes as discussed in Chapters 1 and 3 of Hettmansperger and McKean [7]. For the sign process theory, we assume that the pdf of the random errors is positive at its median. The proof is in two parts. Part 2 gives the desired result, while Part 1 establishes the consistency of the medians used in the second part.

Part 1 of the Proof:

Consider the j-th bin, for $j=1,...M$. Let $\{\mathbf{e}_j\}$ denote the process of random errors of the linear model $\mathbf{Z}_j=\mathbf{X}_j\boldsymbol{\beta}_j+\mathbf{e}_j$. Assume without loss of generality that $\boldsymbol{\beta}_j=\mathbf{0}$ and the median of e_i is 0, where for ease of notation we have omitted the second subscript j on e_i. Let $\hat{\mathbf{e}}=\mathbf{Z}_j-\mathbf{X}_j\hat{\boldsymbol{\beta}}_{j,ls}$ denote the residuals from the a fit such that $\sqrt{n}\,(\hat{\beta}-\beta)=\mathcal{O}(1)$. Let $F(t)$ and $f(t)$ denote the cdf and pdf of e_i, respectively.

Consider the sign process given by

$$\overline{S}_j(\theta)=\frac{1}{n_j}\sum_{i=1}^{n_j}\operatorname{sgn}(e_i-\theta)\,, \tag{A.1}$$

where $\operatorname{sgn}(u)=-1,0,\text{or }1$ for $u<0$, $u=0$, or $u>0$. Denote the median of $e_1,...,e_{n_j}$ by $\hat{\theta}_e$. Notice that $\hat{\theta}_e$ solves the equation $\overline{S}_j(\theta)=0$. Our immediate goal is the asymptotic linearity of the process $\overline{S}_j(\theta)$ that is given in expression (A.3). We accomplish this by showing that the four sufficient conditions hold as given in Section 1.5 of Hettmansperger and McKean [7]. First note that $\overline{S}_j(\theta)$ is a nonincreasing function of θ. Thus the first condition holds. For the second condition, by a simple shift theorem and stationarity, we have

$$\mu(\theta)\,=\,E_0\big[\overline{S}_j(\theta)\big]\,=\,E_\theta\big[\overline{S}_j(0)\big]\,=\,\frac{1}{n_j}\sum_{i=1}^{n_j}E_\theta\big[\operatorname{sgn}(e_i)\big]\,=\,1-2F(-\theta)\,.$$

Hence, $\mu'(0)=2f(0)>0$ which establishes the second condition.

For the third condition, we need to show the variance of $\sqrt{n_j}\,\overline{S}_j(0)$ exists. This variance is

$$\begin{aligned}\sigma^2_{n_j}&=V\big[\sqrt{n_j}\,\overline{S}_j(0)\big]\\&=\frac{1}{n_j}\sum_{i=1}^{n_j}V\big(\operatorname{sgn}(e_i)\big)+\frac{2}{n_j}\sum_{i=1}^{n_j-1}\sum_{k=i+1}^{n_j}\operatorname{cov}\big[\operatorname{sgn}(e_i),\operatorname{sgn}(e_k)\big]\,.\end{aligned}$$

The first term on the right is easily seen to be 1. Using $P[e_i<0]=1/2$ and expanding each covariance term into its expectation, we obtain four probability terms and, hence, the sum of four series. The absolute value of one of these four series is given next. As we show, we establish a bound on the series by invoking the assumption (3.1) and then applying properties of the geometric series. A similar proof holds for the other three series.

$$
\begin{aligned}
\left| \frac{2}{n_j} \sum_{i=1}^{n_j-1} \sum_{k=i+1}^{n_j} \Big[P(e_i < 0,\, e_k < 0) - P(e_i < 0)\, P(e_k < 0) \Big] \right| \leq \\
&\leq \frac{2}{n_j} \sum_{i=1}^{n_j-1} \sum_{k=i+1}^{n_j} \Big| P(e_i < 0,\, e_k < 0) - P(e_i < 0)\, P(e_k < 0) \Big| \\
&\leq K \frac{2}{n_j} \sum_{i=1}^{n_j-1} \sum_{k=i+1}^{n_j} \rho^{k-i} \\
&= 2K \frac{\rho}{1-\rho} \frac{n_j - 1}{n_j} - \left[\frac{1}{n_j} \frac{\rho^2}{1-\rho^2} \left(1 - \rho^{n_j - 1}\right) \right] \\
&\leq 2K \frac{\rho}{1-\rho} \,,
\end{aligned}
$$

where the constants $K > 0$ and $0 \leq \rho < 1$ are given in expression (3.1). The last line follows because the term in brackets is nonnegative and the entire expression is nonnegative. Thus the above series is convergent. Since the other three series follow similarly and, since absolute convergence implies convergence, the series for the variance $\sigma^2_{n_j}$ converges. Let $\sigma^2(0)$ denote the value to which the series converges. The actual value is not needed in the proof but can be obtained from Wendler [20] as noted below.

The fourth condition requires that for all b, $\mathrm{Var}_0\Big\{ \sqrt{n_j}\, \big[\overline{S}(b/\sqrt{n_j}) - \overline{S}(0) \big] \Big\} \to 0$, as $n_j \to \infty$, where $I(x) = 1$ if x is true, 0 otherwise. Based on the sign function, we have

$$
V_{n_j,b} =_{\mathrm{dfn}} \mathrm{Var}\Big[\sqrt{n_j}\, \big[\overline{S}(b/\sqrt{n_j}) - \overline{S}(0) \big] \Big] = \mathrm{Var}\left[\frac{-2}{\sqrt{n_j}} \sum_{i=1}^{n_j} I\big(0 < e_i < b/\sqrt{n_j}\big) \right].
$$

Thus,

$$
\begin{aligned}
V_{n_j,b} &= \frac{4}{n_j} \sum_{i=1}^{n_j} \mathrm{Var}\Big[I\big(0 < e_i < b/\sqrt{n_j}\big) \Big] \\
&\quad + \frac{8}{n_j} \sum_{i=1}^{n_j-1} \sum_{k=i+1}^{n_j} \mathrm{cov}\Big[I\big(0 < e_i < b/\sqrt{n_j}\big), I\big(0 < e_k < b/\sqrt{n_j}\big) \Big].
\end{aligned}
\tag{A.2}
$$

By stationarity and continuity of the cdf $F(t)$, $E\big[I\big(0 < e_i < b/\sqrt{n_j}\big) \big] = F\big(b/\sqrt{n_j}\big) - \frac{1}{2} \to 0$, as $n_j \to \infty$; hence, the variance term on the right side of (A.2) goes to 0 as $n_j \to \infty$.

We can write the covariances as

$$
\begin{aligned}
c_{n_j,i,k} &=_{\mathrm{dfn}} \mathrm{cov}\Big[I\big(0 < e_i < b/\sqrt{n_j}\big), I\big(0 < e_k < b/\sqrt{n_j}\big) \Big] \\
&= P\Big[0 < e_i < b/\sqrt{n_j}\,,\; 0 < e_k < b/\sqrt{n_j} \Big] \\
&\quad - P\Big[0 < e_i < b/\sqrt{n_j} \Big]\, P\Big[0 < e_k < b/\sqrt{n_j} \Big].
\end{aligned}
$$

Notice that this is similar to the above argument on the variance, except that the terms also go to zero as $n_j \to \infty$. Using mean value theorems it follows that the rate of this convergence is $1/n_j$. Using the assumptions from Section 3 and this rate we have $|c_{n_j,i,k}| \leq K \rho^{k-i}_{n_j}$, where $\rho_{n_j} = O(1/n_j)$. Following the same argument as used for the variance, the covariance term in (A.2) in absolute value is less than or equal to

$$
2K \frac{\rho_{n_j}}{1 - \rho_{n_j}} \leq O(1/n_j) \to 0\,, \qquad \text{as} \quad n_j \to \infty\,.
$$

Thus $V_{n_j,b} \to 0$ as $n_j \to \infty$.

By these four conditions, as shown in Chapter 1 of Hettmansperger and McKean [7], the sign process satisfies the linearity result:

$$\sqrt{n_j}\,\overline{S}_j(\theta) \;=\; \sqrt{n_j}\,\overline{S}_j(0) - 2f(0)\sqrt{n_j}\,\theta + o_p(1)\,, \tag{A.3}$$

for $\sqrt{n_j}\,|\theta| \le B$, for all $B > 0$.

To obtain $\sigma^2(0)$, we can use Wendler [20]. He showed, under the mixing conditions above, that $\sqrt{n_j}\,|\hat{\theta}_e|$ converges in distribution and, hence, is tight. Since $\overline{S}_j(\theta) = 0$, we can use (A.3) and Wendler's asymptotic distribution to obtain the asymptotic normal distribution of $\sqrt{n_j}\,\overline{S}_j(0)$.

For our proof, we are interested in the residual process. Since for the proof the true parameters are 0, we can write the residuals as $\hat{e}_i = e_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{LS}$, $i = 1, ..., n_j$. The residual sign process is then given by

$$\overline{S}_j^*(\theta) = \frac{1}{n_j}\sum_{i=1}^{n_j} \operatorname{sgn}(\hat{e}_i - \theta)\,. \tag{A.4}$$

Let $\hat{\theta}^*$ denote median of the residuals. Notice that it solves $\overline{S}_j^*(\hat{\theta}^*) = 0$. In the independent error case, Hettmansperger and McKean [7] established the linearity of the residual process for any root-n consistent estimate of $\boldsymbol{\beta}$; see their Section 3.5 and the associated parts of the Appendix. A key result used in their proof was the linearity for the single sample case, i.e., in the current proof, the result (A.3). See Lemma A.3.2 of Hettmansperger and McKean [7]. The remainder of the proof for the linearity of $\overline{S}_j^*(\theta)$ follows using similar reasoning as above. The result is

$$\sqrt{n_j}\,\overline{S}_j^*(\theta) \;=\; \sqrt{n_j}\,\overline{S}_j^*(0) - 2f(0)\sqrt{n_j}\,\theta + o_p(1)\,, \tag{A.5}$$

for $\sqrt{n_j}\,|\theta| \le B$, for all $B > 0$. Using this and $\overline{S}_j^*(\hat{\theta}^*) = 0$, we obtain the asymptotic distribution of $\hat{\theta}^*$ and, hence, its consistency.

The second part of our proof requires the consistency of three other estimators. The first is the median of the absolute value of the residuals. This is easily obtained by replacing e_i with $|e_i|$ in the above processes. Since the pdf of $|e_i|$ is strictly positive at the true median, the proof holds in this case too. The second estimator is a function of the residuals from two bins, say, j and j'. More specifically, it is a function of the residuals

$$\hat{e}_{j,i} + \hat{e}_{j',i'} \;=\; e_{j,i} + e_{j',i'} - \begin{bmatrix} \mathbf{x}_{j,i}^{\mathsf{T}} & \mathbf{x}_{j',i'}^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}_j \\ \hat{\boldsymbol{\beta}}_{j'} \end{bmatrix},$$

where $\hat{\boldsymbol{\beta}}_j$ and $\hat{\boldsymbol{\beta}}_{j'}$ denote the LS estimates from bins j and j', respectively. Because the vector $(\hat{\boldsymbol{\beta}}_j^{\mathsf{T}}, \hat{\boldsymbol{\beta}}_{j'}^{\mathsf{T}})^{\mathsf{T}}$ is root-n consistent and the convolution of identical pdfs is positive at its median when each pdf is positive at its median, nothing in the above proof precludes the use of random errors of the form $e_{j,i} + e_{j',i'}$. Thus the theory holds in this case also. These comments apply to the third estimator also because it is based on the residuals $\hat{e}_{j,i} - \hat{e}_{j',i'}$.

Part 2 of the Proof:

This part of the proof makes use of the standard inequality $|a| = |a-b+b| \leq |a-b| + |b|$. It suffices to show consistency of $\hat{\mathbf{\Sigma}}_M^{(\text{rob})}$ element-wise. We first show the consistency of the diagonal elements. The statistic and functional of the m^{th} diagonal of $\hat{\mathbf{\Sigma}}_M^{(\text{rob})}$ are given by:

$$\text{MAD}\{\hat{\mathbf{e}}_m\} = \text{med}_i\big|\hat{\mathbf{e}}_{m_i} - \text{med}_j\{\hat{\mathbf{e}}_{m_j}\}\big| \quad \text{with functional} \quad \xi_m = \text{med}\big|\mathbf{e}_m - \text{med}\{\mathbf{e}_m\}\big|\,.$$

Without loss of generality, assume that $\text{med}\{\mathbf{e}_m\} = 0$. From Part 1, $\text{med}_i\{\hat{\mathbf{e}}_{m_i}\} \xrightarrow{P} 0$, in probability. Next, assume that $\text{med}\{|\mathbf{e}_m|\} = \xi$. Then also from Part 1, $\text{med}_i|\hat{\mathbf{e}}_{m_i}| \xrightarrow{P} \xi$. Choose N_0 sufficiently large so that, given $\varepsilon > 0$,

$$k \geq N_0 \implies \big|\text{med}_{1\leq i\leq k}\{\hat{\mathbf{e}}_{m_i}\}\big| < \varepsilon \tag{A.6}$$

with probability greater than $(1-(\varepsilon/2))$. Let A_n denote the event where (A.6) occurs. Then, on A_n we have

$$\begin{aligned} |\hat{\mathbf{e}}_{m_i}| &= \big|\hat{\mathbf{e}}_{m_i} - \text{med}_j\{\hat{\mathbf{e}}_{m_j}\} + \text{med}_j\{\hat{\mathbf{e}}_{m_j}\}\big| \\ &\leq \big|\hat{\mathbf{e}}_{m_i} - \text{med}_j\{\hat{\mathbf{e}}_{m_j}\}\big| + \big|\text{med}_j\{\hat{\mathbf{e}}_{m_j}\}\big| \\ &< \big|\hat{\mathbf{e}}_{m_i} - \text{med}_j\{\hat{\mathbf{e}}_{m_j}\}\big| + \varepsilon\,. \end{aligned}$$

So, on A_n,

$$\text{med}_i|\hat{\mathbf{e}}_{m_i}| < \text{med}_i\big|\hat{\mathbf{e}}_{m_i} - \text{med}_j\{\hat{\mathbf{e}}_{m_j}\}\big| + \varepsilon\,, \tag{A.7}$$

and

$$\begin{aligned} \big|\hat{\mathbf{e}}_{m_i} - \text{med}_j\{\hat{\mathbf{e}}_{m_j}\}\big| &= \big|\hat{\mathbf{e}}_{m_i} - \text{med}_j\{\hat{\mathbf{e}}_{m_j}\} - \hat{\mathbf{e}}_{m_i} + \hat{\mathbf{e}}_{m_i}\big| \\ &\leq \big|\text{med}_j\{\hat{\mathbf{e}}_{m_j}\}\big| + |\hat{\mathbf{e}}_{m_i}| \\ &< |\hat{\mathbf{e}}_{m_i}| + \varepsilon\,. \end{aligned}$$

Hence, on A_n,

$$\text{med}_i\big|\hat{\mathbf{e}}_{m_i} - \text{med}_j\{\hat{\mathbf{e}}_{m_j}\}\big| < \text{med}_i|\hat{\mathbf{e}}_{m_j}| + \varepsilon\,. \tag{A.8}$$

Putting (A.7) and (A.8) together, we have on A_n,

$$\Big|\text{med}_i\big|\hat{\mathbf{e}}_{m_i} - \text{med}_j\{\hat{\mathbf{e}}_{m_j}\}\big| - \text{med}_i|\hat{\mathbf{e}}_{m_i}|\Big| < \varepsilon\,. \tag{A.9}$$

Since this occurs with probability of at least $(1-(\varepsilon/2))$, the difference on the left-side goes to 0 in probability. As noted above, from Part 1, $\text{med}_i|\hat{\mathbf{e}}_{m_i}| \xrightarrow{P} \xi$; hence, $\text{med}_i\big|\hat{\mathbf{e}}_{m_i} - \text{med}_j\{\hat{\mathbf{e}}_{m_j}\}\big| \xrightarrow{P} \xi$.

For the off-diagonal elements, let $m \neq m'$ be given. Recall that the off-diagonal elements of $\hat{\mathbf{\Sigma}}_M^{(\text{rob})}$ are given by equation (2.2), which can be expressed as follows:

$$\hat{\mathbf{\Sigma}}_M^{(\text{rob})}(m, m') = \left(\text{MAD}\left\{\frac{\hat{\mathbf{e}}_m \oplus \hat{\mathbf{e}}_{m'}}{2}\right\}\right)^2 - \left(\text{MAD}\left\{\frac{\hat{\mathbf{e}}_m \ominus \hat{\mathbf{e}}_{m'}}{2}\right\}\right)^2. \tag{A.10}$$

It suffices to show consistency for each of the terms on the right-side. Define $\mathbf{t} = \frac{1}{2}\left(\mathbf{e}_m \oplus \mathbf{e}_{m'}\right)$. Then the statistic and its functional, respectively, for the off-diagonal elements are:

$$\text{MAD}\{\hat{\mathbf{t}}\} = \text{med}_i\big|\hat{\mathbf{t}}_i - \text{med}_j\{\hat{\mathbf{t}}_j\}\big| \quad \text{with functional} \quad \xi_{m,m'} = \text{med}\big|\mathbf{t} - \text{med}\{\mathbf{t}\}\big|\,.$$

Without loss of generality let $\text{med}\{\mathbf{t}\} = 0$. From Part 1, $\text{med}_i\{\hat{\mathbf{t}}_i\} \xrightarrow{P} 0$. Then the proof follows in the same manner as for the diagonal elements. So each of the MADs in equation (A.10) is consistent. Therefore, the entire expression is consistent. Thus, the diagonal and off-diagonal entries of $\hat{\mathbf{\Sigma}}_M^{(\text{rob})}$ are consistent. Hence, $\hat{\mathbf{\Sigma}}_M^{(\text{rob})}$ is a consistent estimator of $\mathbf{\Sigma}_M$. □

ACKNOWLEDGMENTS

We thank the editors of *REVSTAT* as well as the reviewers, which provided valuable feedback and suggestions to improve this manuscript.

The methods described in this manuscript have been implemented in the `R` programming language. The code is available at https://github.com/jelsema/RRSM. Should the link deprecate, contact the first author for information regarding how to obtain the code.

REFERENCES

[1] Banerjee, S.; Gelfand, A.; Finley, A. and Sang, H. (2008). Gaussian predictive process models for large spatial datasets, *Journal of the Royal Statistical Society: Series B*, **70**(4), 825–844.

[2] Bliznyuk, N.; Carroll, R.; Genton, M. and Wang, Y. (2012). Variogram estimation in the presence of trend, *Statistics and Its Interface*, **5**, 159–168.

[3] Cressie, N. and Hawkins, D. (1984). Robust kriging – a proposal, *Mathematical Geology*, **16**, 3–18.

[4] Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**, 209–226.

[5] Finley, A.O.; Sang, H.; Banerjee, S. and Gelfand, A.E. (2009). Improving the performance of predictive process modeling for large datasets, *Computational Statistics & Data Analysis*, **53**(8), 2873–2884.

[6] French, J. (2018). *SpatialTools: Tools for Spatial Data Analysis*, R package version 1.0.4.

[7] Hettmansperger, T. and McKean, J. (2011). *Robust Nonparametric Statistical Methods*, Chapman Hall, New York, 2nd edition.

[8] Kang, E. and Cressie, N. (2011). Bayesian inference for the spatial random effects model, *Journal of the American Statistical Association*, **106**, 972–983.

[9] Kang, E.; Cressie, N. and Shi, T. (2010). Using temporal variability to improve spatial mapping with application to satellite data, *The Canadian Journal of Statistics*, **38**, 271–289.

[10] Kang, E.L.; Cressie, N. and Sain, S.R. (2012). Combining outputs from the north american regional climate change assessment program by using a bayesian hierarchical model, *Journal of the Royal Statistical Society C*, **61**(2), 291–313.

[11] Katzfuss, M. and Cressie, N. (2011). *Tutorial on fixed rank kriging (frk) of co2 data*, Technical Report, The Ohio State University, 858.

[12] Koenker, R. (2018). *quantreg: Quantile Regression*, R package version 5.35.

[13] Koenker, R. and Bassett, G. (1978). Regression quantiles, *Econormetrica*, **46**, 33–50.

[14] Lahiri, S.; Lee, Y. and Cressie, N. (2002). Asymptotic distribution and asymptotic efficiency of least squares estimators of variogram parameters, *Journal of Statistical Planning and Inference*, **103**, 65–85.

[15] Nychka, D.; Wikle, C. and Royle, J.A. (2002). Multiresolution models for nonstationary spatial covariance functions, *Statistical Modelling*, **2**, 315–331.

[16] Paul, R.; Jelsema, C.M. and Lau, K.W. (2015). *A flexible class of reduced rank spatial models for large non-gaussian datasets.* In "Current Trends in Bayesian Methodology with Applications" (S.K. Upadhyay, U. Singh, D.K. Dey and A. Loganathan, Eds.), Chapman & Hall/CRC Press.

[17] Sang, H. and Huang, J.Z. (2011). A full scale approximation of covariance functions for large spatial data sets, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **74**(1), 111–132.

[18] Shi, T. and Cressie, N. (2007). Global statistical analysis of misr aerosol data: a massive data product from nasa's terra satellite, *Environmetrics*, **18**, 665–680.

[19] Stein, M.L. (2007). Spatial variation of totla column ozone on a global scale, *The Annals of Applied Statistics*, **1**, 191–210.

[20] Wendler, M. (2011). Bahadur representation for U-quantiles of dependent data, *Journal of Multivariate Analysis*, **102**, 1064–1079.

[21] Wilks, D. (2006). *Statistical Methods in the Atmospheric Sciences*, Academic Press, San Diego, USA, 2nd edition.

[22] Zhu, Y.; Kang, E.L.; Bo, Y.; Tang, Q.; Cheng, J. and He, Y. (2015). A robust fixed rank kriging method for improving the spatial completeness and accuracy of satellite sst products, *IEEE Transactions on Geoscience and Remote Sensing*, **53**, 5021–5035.