Optimal Robust Methods for Prediction and Design in Spatial Studies

Running title: Spatial Prediction and Design

Douglas P. Wiens[1]

University of Alberta

September 22, 2003

This report was subsequently split into two papers, entitled *Robustness in Spatial Studies I: Minimax Estimation and Prediction* and *Robustness in Spatial Studies II: Minimax Design*, published in *Environmetrics*. In the process the neighbourhood structure was changed.

We develop and test robust methods for design construction, for estimation and for prediction in spatial studies. The designs are robust against misspecified variance/covariance structures, and against misspecified regression responses. Robustness against contaminated error distributions is provided by the use of generalized M-estimators in the estimation and prediction procedures. The loss function is based on the mean squared error of the predicted values. This is maximized, analytically, over the various neighbourhoods quantifying the departures from the fitted model. This maximum is then minimized numerically - by simulated annealing, or sequentially - in order to obtain the optimal designs.

AMS 2000 SUBJECT CLASSIFICATIONS: Primary 62K05, 62F35, 62M30; secondary 62L05, 62G35.

KEY WORDS: Breakdown; Environmental monitoring; Generalized M-estimation; Isotropic; Minimax; M-estimate; Sequential; Simulated annealing.

---

[1]Douglas P. Wiens is Professor, Statistics Centre, Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2G1. Tel.: (780) 492-4406; fax.: (780) 492-4826; e-mail: doug.wiens@ualberta.ca.

1. INTRODUCTION

Consider the following problem, of interest in environmetrics. There is a set $\mathcal{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_N\}$ of locations at which environmental monitoring stations may be placed. The agency responsible is to choose, from these, locations $\mathcal{S} = \{\mathbf{t}_{i_1}, \ldots, \mathbf{t}_{i_n}\}$. At these locations one will observe, with measurement error, a stochastic process $X(\mathbf{t})$ (air quality index, pollution level, etc.): $Y(\mathbf{t}) = X(\mathbf{t}) + \epsilon(\mathbf{t})$. The purpose is to predict $X(\mathbf{t})$, possibly for $\mathbf{t} \in \mathcal{T} \setminus \mathcal{S}$. Suppose however that - as is typically the case in spatial studies - the spatial correlation structure of $\{X(\mathbf{t}) \,|\mathbf{t} \in \mathcal{T}\}$ is possibly misspecified, that there is a possibly misspecified regression structure governing $E[X(\mathbf{t})]$, and that $\epsilon(\mathbf{t})$, while uncorrelated with $X(\mathbf{t})$ and with $\epsilon(\mathbf{t}')$ for $\mathbf{t} \neq \mathbf{t}'$, has a possibly misspecified variance structure. In the face of these model uncertainties, one is to choose $\mathcal{S}$ in some "optimal" manner, and then do (robust) estimation and prediction.

The above is an outline of the problems addressed in this article, the sole generalization being that no restriction is made on the dimension of $\mathbf{t} \in \mathbb{R}^d$. Our framework is sufficiently broad as to encompass several scenarios:

1. No locations have yet been chosen, and we are free to choose any n sites.

2. There is an existing network of n_0 sites at locations $\mathcal{S}_0$ and we are to choose $n - n_0$ further sites. Scenario 1 is this case with $n_0 = 0$. See Thompson (1997, p. 18) for a discussion.

3. There is an existing network of n_1 sites at locations $\mathcal{S}_1$ and we must *eliminate* $n_1 - n$ of them. This is equivalent to setting $\mathcal{T} = \mathcal{S}_1$ and then finding the n best sites which are to remain.

We assume that:

- $\mathrm{VAR}[\epsilon(\mathbf{t})] = f^2(\mathbf{t})$ for a variance function $f^2(\cdot)$. Define $\mathbf{F}_{N\times N} = diag\left(f^2(\mathbf{t}_1), \ldots, f^2\left(\mathbf{t}_N\right)\right)$.
- $X(\mathbf{t}) = E[X(\mathbf{t})] + \delta(\mathbf{t})$, where $\mathrm{COV}[\delta(\mathbf{t}), \delta(\mathbf{t}')] = g(\mathbf{t}, \mathbf{t}')$ for a covariance function g. Define $\mathbf{G}_{N\times N}$ by $g_{ij} = g\left(\mathbf{t}_i, \mathbf{t}_j\right)$.

- The mean response is approximately linear in regressors $z_j(\mathbf{t})$:

$$E\left[X(\mathbf{t})\right] \approx \mathbf{z}^T(\mathbf{t})\boldsymbol{\theta},$$

where $\mathbf{z}(\mathbf{t}) = (z_1(\mathbf{t}), ..., z_p(\mathbf{t}))^T$ and the parameter vector $\boldsymbol{\theta}_{p\times 1}$ makes the approximation most accurate, *viz.*,

$$\boldsymbol{\theta}\ (=\boldsymbol{\theta}_N) = \arg\min_{\mathbf{v}} \sum_{\mathbf{t}\in\mathcal{T}} \left(E\left[X(\mathbf{t})\right] - \mathbf{z}^T(\mathbf{t})\mathbf{v}\right)^2. \tag{1}$$

We define $h(\mathbf{t})$ such that

$$E\left[X(\mathbf{t})\right] = \mathbf{z}^T(\mathbf{t})\boldsymbol{\theta} + h(\mathbf{t}).$$

The definition of $\boldsymbol{\theta}$ implies the orthogonality condition

$$\sum_{\mathbf{t}\in\mathcal{T}} \mathbf{z}(\mathbf{t})h(\mathbf{t}) = \mathbf{0}_{p\times 1}. \tag{2}$$

We aim to predict a set $\mathbf{Cx}$ of M linear functions of $\mathbf{x} = (X(\mathbf{t}_1), \ldots, X(\mathbf{t}_N))^T$. It is our intention to obtain predictors which are robust against misspecified functions f, g and h. For example, f could be erroneously specified as constant, g erroneously specified as isotropic, h erroneously specified as $\equiv 0$.

Spatial design problems for correctly specified models have been studied by Martin (1986), Fedorov and Hackl (1994), Stein (1995) and Thompson (1997), among others. Schilling (1992) and McArthur (1987) assess some particular sampling designs. To our knowledge this is the first work to explicitly seek robustness of spatial design against model uncertainties.

We will proceed as follows. Let $\mathbf{y} = (Y(\mathbf{t}_{i_1}), \ldots, Y(\mathbf{t}_{i_n}))^T$. For a matrix $\mathbf{A}_{M\times n}$ defining a set $\mathbf{Ay}$ of linear predictors, the mean squared error matrix is

$$\mathbf{MSE}\left(\mathbf{A}; f, g, h\right) = E\left[\left(\mathbf{Ay} - \mathbf{Cx}\right)\left(\mathbf{Ay} - \mathbf{Cx}\right)^T\right].$$

In §2 we derive the best linear unbiased predictors $\mathbf{A}_0\mathbf{y}$, assuming that f, g and h are correctly specified as f_0, g_0 and $h_0 \equiv 0$. Our loss function is the average diagonal element of the *mse* matrix:

$$\mathcal{L}\left(\mathbf{A}; f, g, h\right) = M^{-1} tr\left(\mathbf{MSE}\left(\mathbf{A}; f, g, h\right)\right). \tag{3}$$

Thus

$$\mathbf{A}_0 = \arg\min_{\mathbf{A}\in\mathcal{A}} \mathcal{L}\left(\mathbf{A}; f_0, g_0, h_0\right), \tag{4}$$

where $\mathcal{A}$ is the class of $M \times n$ matrices satisfying the unbiasedness constraint

$$E[\mathbf{A}\mathbf{y}] = E[\mathbf{C}\mathbf{x}] \text{ for all } \boldsymbol{\theta}. \tag{5}$$

The corresponding estimate of $\boldsymbol{\theta}$ is the Generalized Least Squares Estimate (GLSE). We shall also discuss a "robustified" predictor, optimized for use with a Generalized M-estimate (GM-estimate) of $\boldsymbol{\theta}$.

We exhibit the "realized" loss $\mathcal{L}(\mathbf{A}_0; f, g, h)$, attained when $\mathbf{A}_0$ is used but the true functions are f, g and h. This sets the stage for a treatment, in §3, of our robust design problems - we aim to choose the locations $\mathcal{S}$ so as to minimize the maximum realized mean squared error, as f, g and h range over neighbourhoods of f_0, g_0 and h_0 defined by

$$\begin{aligned}
\mathcal{F}_\alpha &= \left\{ f(\cdot) \mid tr\,(\mathbf{F} - \mathbf{F}_0)^2 \le N\alpha^2,\ f(\mathbf{t}) \ge 0 \right\}, \\
\mathcal{G}_\beta &= \left\{ g(\cdot,\cdot) \mid tr\,(\mathbf{G} - \mathbf{G}_0)^2 \le N\beta^2,\ \mathbf{G} \ge \mathbf{0} \right\}, \\
\mathcal{H}_\gamma &= \left\{ h(\cdot) \mid \sum_{\mathbf{t}\in\mathcal{T}} h^2(\mathbf{t}) \le \sqrt{N}\gamma,\ \sum_{\mathbf{t}\in\mathcal{T}} \mathbf{z}(\mathbf{t}) h(\mathbf{t}) = \mathbf{0} \right\}.
\end{aligned}$$

In these definitions $\mathbf{F}_0$ and $\mathbf{G}_0$ are the covariance matrices under f_0 and g_0 and $\mathbf{G} \ge \mathbf{0}$ refers to the ordering by non-negative definiteness. We specify a smaller radius for $\mathcal{H}_\gamma$ than for $\mathcal{F}_\alpha$ or $\mathcal{G}_\beta$ in order that the contributions of bias and variance to mean squared error be approximately of the same magnitude.

A mathematical description of our design problem is that we design so as to minimize $\max_{f,g,h} \mathcal{L}(\mathbf{A}_0; f, g, h)$ where, as will be seen below, $\mathbf{A}_0$ defined by (4) results in the use of the well known "universal kriging" predictor, or a robustification of this predictor as described in §2.2. A possible alternate approach would be to instead minimize $\max_{f,g,h} \mathcal{L}(\mathbf{A}; f, g, h)$ over both the design and $\mathbf{A}$, i.e. we might also seek a minimax linear predictor. There is some precedent for such an approach - Marcus and Sacks (1976), Heckman (1987) - in design problems in which parameter estimation is the primary goal. The optimal linear estimator is then typically not the classical GLSE. Given the popularity of the GLSE among linear estimates, we have opted to proceed as described above.

Example 1.1: If $p = 1$ and $z(\mathbf{t}) = 1$ then $X(\mathbf{t})$ has "approximately constant" mean $\theta + h(\mathbf{t})$, where $\sum_{\mathbf{t}\in\mathcal{T}} h(\mathbf{t}) = 0$. Then if $M = 1$ and $\mathbf{C} = (1, .., 1)$ we are predicting $X_{Total} = \sum_{\mathbf{t}\in\mathcal{T}} X(\mathbf{t})$ by a linear function $\hat{X}_{Total} = \mathbf{a}^T\mathbf{y}$. The loss in this case is $\mathcal{L}(\mathbf{a}; f, g, h) = E\left[\left(\hat{X}_{Total} - X_{Total}\right)^2\right]$.

Example 1.2: Suppose that $M = N - n$ and $\mathbf{C}$ is the incidence matrix for $\mathcal{T} \backslash \mathcal{S}$, i.e. $\mathbf{C}$ is the result of omitting, from $\mathbf{I}_N$, rows $i_1, ..., i_n$. Then we are predicting $X(\mathbf{t})$ by a linear function $\hat{X}(\mathbf{t}) = \mathbf{a}_\mathbf{t}^T \mathbf{y}$ for each $\mathbf{t} \notin \mathcal{S}$. The matrix $\mathbf{A}$ has rows $\mathbf{a}_\mathbf{t}^T$, and the loss is

$$\frac{1}{N-n} \sum_{\mathbf{t} \notin \mathcal{S}} E\left[\left(\hat{X}(\mathbf{t}) - X(\mathbf{t})\right)^2\right]. \tag{6}$$

If instead $M = N$ and $\mathbf{C} = \mathbf{I}_N$, then the loss is

$$\frac{1}{N} \sum_{\mathbf{t} \in \mathcal{T}} E\left[\left(\hat{X}(\mathbf{t}) - X(\mathbf{t})\right)^2\right], \tag{7}$$

which is the Average Mean Squared Prediction Error (AMSPE). Our robust optimality criterion is then analogous to the classical notion of I-optimality.

In a similar vein Sacks and Schiller (1988) considered the construction of designs, assuming that $f_0(\cdot)$ and $g_0(\cdot, \cdot)$ were correctly specified and that $E[X(\mathbf{t})]$ was known and $\equiv 0$. They used the loss function $\max_\mathbf{t} E\left[\left(\mathbf{a}_\mathbf{t}^T \mathbf{y} - X(\mathbf{t})\right)^2\right]$, and remarked upon the lack of robustness, to changes in g_0, of their procedures.

Example 1.3: If $p = d + 1$ and $\mathbf{z}(\mathbf{t}) = \left(1, \mathbf{t}^T\right)^T$ then the regression response is approximately linear in the coordinates of $\mathbf{t}$.

The maximum loss is exhibited in Theorem 2, §3. In §4 we obtain optimal robust designs by minimizing this maximum loss. For small values of N and n this can be done exactly, by performing an exhaustive search of all $\binom{N}{n}$ designs. For more realistic values we investigate two algorithms, both of which seem to give at least nearly optimal solutions in reasonable amounts of time. The first is a simulated annealing algorithm, whereas the second employs a sequential search technique.

Post-design, robust parameter estimation and process prediction are considered in §5, where we also carry give the results of a simulation study. A summary of our findings is that considerable benefits are to be gained by robust estimation and prediction procedures, when the error distribution is contaminated, for only a small premium in efficiency at the assumed model. Both robust and nonrobust procedures gain in efficiency when combined with a judiciously chosen design. In §6 we revisit a data set from the literature, in order to illustrate an application of our methods.

All derivations are in the Appendix.

2. OPTIMAL AND ROBUST PREDICTION

In this section we shall first determine the optimal *linear* predictor of $\mathbf{Cx}$, assuming that $\mathbf{F}_0$ and $\mathbf{G}_0$ are completely and correctly specified and that $h_0 \equiv 0$. We exhibit the loss $\mathcal{L}(\mathbf{A}_0; f, g, h)$, attained at arbitrary $f \in \mathcal{F}_\alpha, g \in \mathcal{G}_\beta, h \in \mathcal{H}_\gamma$. We then discuss modifications resulting from replacing, in the predictor, the GLSE $\hat{\boldsymbol{\theta}}$ by a Generalized M-estimate.

The observed data vector $\mathbf{y}$ may be decomposed as $\mathbf{y} = \mathbf{Q}_1(\mathbf{x} + \boldsymbol{\varepsilon})$, where $\mathbf{Q}_1 : n \times N$ is the incidence matrix for $\mathcal{S}$, $\mathbf{x} = (X(\mathbf{t}_1), ..., X(\mathbf{t}_N))^T$ and $\boldsymbol{\varepsilon} = (\varepsilon(\mathbf{t}_1), ..., \varepsilon(\mathbf{t}_N))^T$. Define

$$\mathbf{h} = (h(\mathbf{t}_1), ..., h(\mathbf{t}_N))^T, \quad \mathbf{Z} = (\mathbf{z}(\mathbf{t}_1), ..., \mathbf{z}(\mathbf{t}_N))^T,$$

$$\mathbf{Z}_1 = \mathbf{Q}_1\mathbf{Z} : n \times p, \qquad \mathbf{G}_1 = \mathbf{Q}_1\mathbf{G} : n \times N,$$

$$\mathbf{F}_{11} = \mathbf{Q}_1\mathbf{F}\mathbf{Q}_1^T : n \times n, \qquad \boldsymbol{\Sigma}_{11} = \mathbf{Q}_1(\mathbf{G} + \mathbf{F})\mathbf{Q}_1^T : n \times n.$$

We assume throughout that $\boldsymbol{\Sigma}_{11}$ is positive definite.

In this notation the unbiasedness condition (5) is equivalent to

$$\mathbf{AZ}_1 - \mathbf{CZ} = (\mathbf{AQ}_1 - \mathbf{C})\mathbf{Z} = \mathbf{0}_{M\times p}, \tag{8}$$

and the generalized least squares estimator (GLSE) of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}} = \mathbf{Ry}$, where

$$\mathbf{R} = \left(\mathbf{Z}_1^T\boldsymbol{\Sigma}_{11}^{-1}\mathbf{Z}_1\right)^{-1}\mathbf{Z}_1^T\boldsymbol{\Sigma}_{11}^{-1} : p \times n.$$

2.1. Linear Prediction; $\mathbf{F}_0$ and $\mathbf{G}_0$ completely specified

Theorem 1. *The best linear unbiased predictor (BLUP) of* $\mathbf{Cx}$ *is* $(\widehat{\mathbf{Cx}})_{GLS} = \mathbf{A}_0\mathbf{y}$, *where* $\mathbf{A}_0 = \mathbf{CP}_0 : M \times n$ *for*

$$\mathbf{P}_0 = \mathbf{ZR}_0 + \mathbf{G}_{1,0}^T\boldsymbol{\Sigma}_{11,0}^{-1}(\mathbf{I}_n - \mathbf{Z}_1\mathbf{R}_0) : N \times n. \tag{9}$$

The subscript 0 *indicates evaluation at* $(f, g) = (f_0, g_0)$.
Define the $M \times N$ *matrix* $\mathbf{B}_0 = \mathbf{C}(\mathbf{P}_0\mathbf{Q}_1 - \mathbf{I}_N)$, *with* $\mathbf{B}_0\mathbf{Z} = \mathbf{0}_{M\times p}$. *Let* $\mathbf{d}_0$ *and* $\mathbf{f}_1$ *be the* $n \times 1$ *vectors consisting of the diagonal elements of* $\mathbf{A}_0^T\mathbf{A}_0$ *and* $\mathbf{F}_{11}$ *respectively. For the BLUP the loss (3) is*

$$\mathcal{L}(\mathbf{A}_0; f, g, h) = \left(\mathbf{d}_0^T\mathbf{f}_1 + tr\left(\mathbf{B}_0\mathbf{G}\mathbf{B}_0^T\right) + \|\mathbf{B}_0\mathbf{h}\|^2\right)/M. \tag{10}$$

Remarks:

1. A more succinct description of the BLUP is

$$\widehat{\mathbf{Cx}} = \mathbf{C}\left(\hat{\mathbf{x}} + \mathbf{G}_{1,0}^T \boldsymbol{\Sigma}_{11,0}^{-1}\mathbf{e}\right), \tag{11}$$

where $\hat{\mathbf{x}} = \mathbf{Z}\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{y}} = \mathbf{Z}_1\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\theta}}$ is the GLSE and $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ is the residual vector. If $h \equiv 0$ then the first component $\mathbf{C}\hat{\mathbf{x}}$ is an unbiased estimate of $\mathbf{C}E[\mathbf{x}] = \mathbf{C}\mathbf{Z}\boldsymbol{\theta}$ and the second component has an expected value of $\mathbf{0}$. If as well $f = f_0$ and $g = g_0$, the two components are uncorrelated.

2. If $p = 1$ and $z(\mathbf{t}) = 1$, so that $\mathbf{Z} = \mathbf{1}_N$ and $\mathbf{Z}_1 = \mathbf{1}_n$, then

$$\mathbf{P}_0 = \mathbf{G}_{1,0}^T \boldsymbol{\Sigma}_{11,0}^{-1}\left(\mathbf{I}_n - \frac{\mathbf{1}_n\mathbf{1}_n^T\boldsymbol{\Sigma}_{11,0}^{-1}}{\mathbf{1}_n^T\boldsymbol{\Sigma}_{11,0}^{-1}\mathbf{1}_n}\right) + \frac{\mathbf{1}_N\mathbf{1}_n^T\boldsymbol{\Sigma}_{11,0}^{-1}}{\mathbf{1}_n^T\boldsymbol{\Sigma}_{11,0}^{-1}\mathbf{1}_n}.$$

3. The matrix $\mathbf{B}_0 = \tilde{\mathbf{A}}_0 - \mathbf{C}$, where $\tilde{\mathbf{A}}_0$ is obtained by augmenting $\mathbf{A}_0$ by zeros in columns corresponding to locations in $\mathcal{T} \setminus \mathcal{S}$.

2.2. Generalized M-estimation; $\mathbf{F}_0$ and $\mathbf{G}_0$ completely specified

In this section we discuss generalized M-estimation in the spatial model, and propose a correspondingly robust predictor. Write the regression model as

$$\mathbf{y} = \mathbf{Z}_1\boldsymbol{\theta} + \mathbf{Q}_1\mathbf{h} + \mathbf{Q}_1(\boldsymbol{\varepsilon} + \boldsymbol{\delta}).$$

Let σ_n be a scale functional for $\boldsymbol{\varepsilon}$ such as $(n^{-1}tr\mathbf{F}_{11,0})^{1/2}$, and put $\mathbf{S}_{11,0} = \boldsymbol{\Sigma}_{11,0}/\sigma_n^2$. Then the above becomes

$$\mathbf{v} = \mathbf{U}_{1,0}\boldsymbol{\theta} + \mathbf{k}_1 + \boldsymbol{\eta},$$

where $\mathbf{v} = \mathbf{S}_{11,0}^{-1/2}\mathbf{y}$, $\mathbf{U}_{1,0} = \mathbf{S}_{11,0}^{-1/2}\mathbf{Z}_1$, $\mathbf{k}_1 = \mathbf{S}_{11,0}^{-1/2}\mathbf{Q}_1\mathbf{h}$ and where $\boldsymbol{\eta} = \mathbf{S}_{11,0}^{-1/2}\mathbf{Q}_1(\boldsymbol{\varepsilon} + \boldsymbol{\delta})$ has covariance matrix $\sigma_n^2\mathbf{I}_n$ if $(f, g) = (f_0, g_0)$. Here $\mathbf{S}_{11,0}^{1/2}$ is any matrix satisfying $\mathbf{S}_{11,0}^{1/2}\mathbf{S}_{11,0}^{1/2^T} = \mathbf{S}_{11,0}$. Denote by $\left\{\mathbf{u}_i^T\right\}_{i=1}^n$ the rows of $\mathbf{U}_{1,0}$.

For absolutely continuous, bounded, odd, weakly increasing functions ψ_i and a continuous, bounded, even function χ, the GM-estimate $\hat{\boldsymbol{\theta}}_{GM}$ and corresponding scale estimate $\hat{\sigma}_n$

are defined as members of any sequence satisfying

$$\sum_{i=1}^{n} \psi_i \left(\frac{\hat{\eta}_i}{\hat{\sigma}_n} \right) \mathbf{u}_i = o_P\left(n^{1/2}\right), \tag{12}$$

$$\sum_{i=1}^{n} \left[\chi \left(\frac{\hat{\eta}_i}{\hat{\sigma}_n} \right) - \tau_n \right] = o_P\left(n^{1/2}\right), \tag{13}$$

for $\hat{\eta}_i = v_i - \mathbf{u}_i^T \hat{\boldsymbol{\theta}}_{GM}$ and a bounded sequence of constants $\{\tau_n\}$. For consistency at the normal distribution we take $\tau_n = E_\Phi\left[\chi\left(\eta/\sigma_n\right)\right]$. Common choices of ψ_i are $\psi_i(r) = w(\mathbf{u}_i)\,\psi\left(r/s(\mathbf{u}_i)\right)$ for positive function $w(\cdot), s(\cdot)$. With $s(\mathbf{u}) \equiv 1$ this describes a Mallows-type GM estimate (Hill, 1977). Schweppe (Merrill and Schweppe, 1971) proposed $s(\mathbf{u}) \equiv w(\mathbf{u})$. Corresponding to these proposals, one-step estimates for exactly linear models ($\mathbf{k}_1 \equiv \mathbf{0}$) were investigated by Simpson, Ruppert and Carroll (1992) and Coakley and Hettmansperger (1993). For ordinary M-estimators ($s(\mathbf{u}) \equiv w(\mathbf{u}) \equiv 1,\ \psi_i \equiv \psi$), Silvapullé proved asymptotic normality for fixed (by design) carriers in exactly linear models; Wiens (1996) extended these result to GM-estimators, and to the kinds of approximately linear models considered in the current work. See Field and Wiens (1994) for a study of one-step ordinary M-estimators of regression, under dependence.

To robustify the predictor, we make the observation that (11) can be obtained as the solution to the problem of minimizing, over $M \times n$ matrices $\mathbf{V}$, the MSE of the unbiased predictor $\mathbf{C}\hat{\mathbf{x}} + \mathbf{V}\mathbf{e}$, assuming that $(f, g, h) = (f_0, g_0, 0)$ and that $\hat{\boldsymbol{\theta}}$ is the GLSE. This is because the first order conditions

$$\frac{\partial}{\partial \mathbf{V}} E\left[\left\|\mathbf{C}\left(\hat{\mathbf{x}} - \mathbf{x}\right) + \mathbf{V}\mathbf{e}\right\|^2\right] = \mathbf{0}$$

result in the equations

$$\mathbf{C}E\left[\left(\hat{\mathbf{x}} - \mathbf{x}\right)\mathbf{e}^T\right] + \mathbf{V}E\left[\mathbf{e}\mathbf{e}^T\right] = \mathbf{0}_{M\times n}. \tag{14}$$

For the GLSE these have a solution $\mathbf{V}_{GLS} = \mathbf{C}\mathbf{G}_{1,0}^T \boldsymbol{\Sigma}_{11,0}^{-1}$. To this may be added an arbitrary $M \times n$ member of the row space of $\mathbf{Z}_1^T \boldsymbol{\Sigma}_{11,0}^{-1}$, but since such a matrix is orthogonal to $\mathbf{e}$ it does not affect the predictor.

We propose to replace the GLSE in (11) by $\hat{\boldsymbol{\theta}}_{GM}$, the residuals $\mathbf{e} = \mathbf{S}_{11,0}^{1/2}\left(\mathbf{v} - \mathbf{U}_{1,0}\hat{\boldsymbol{\theta}}\right)$ by robustified residuals $\mathbf{S}_{11,0}^{1/2}\hat{\mathbf{p}}$, where $\hat{\mathbf{p}}$ has elements $\hat{\sigma}_n \psi_i\left(\hat{\eta}_i/\hat{\sigma}_n\right)$, and to obtain an approximate

solution $\mathbf{V}_{GM}$ to (14). As detailed in the Appendix, this results in the optimal robust predictor

$$(\widehat{\mathbf{Cx}})_{GM} = (\widehat{\mathbf{Cx}})_{GLS.rob} + \mathbf{V}_{GLS}\mathbf{S}_{11,0}^{1/2}\mathbf{K}\hat{\mathbf{p}}, \tag{15}$$

where $\mathbf{V}_{GLS} = \mathbf{C}\mathbf{G}_{1,0}^T\left(\sigma_n^2\mathbf{S}_{11,0}\right)^{-1}$, $\mathbf{K} = diag\left(..., \left(\psi_i'(0)/E\left[\psi_i^2\left(\frac{\eta_i}{\sigma_n}\right)\right]\right) - 1, ...\right)$, and

$$(\widehat{\mathbf{Cx}})_{GLS.rob} = \mathbf{C}\left[\mathbf{Z}\hat{\boldsymbol{\theta}}_{GM} + \sigma_n^{-2}\mathbf{G}_{1,0}^T\mathbf{S}_{11,0}^{-1/2^T}\hat{\mathbf{p}}\right]$$

is (11) after making the replacements described above.

Remarks:

4. Simulation studies indicate that the second summand in (15) contributes very little, and can safely be ignored.

5. For the GLSE we have $\mathbf{K} = \mathbf{0}$, and (15) agrees with (11).

6. For a Huber M-estimate $\hat{\boldsymbol{\theta}}_H$ with $\psi_i(r) = \psi(r)$, we have $\mathbf{K} = \kappa\mathbf{I}_n$,where

$$\kappa = \left(\psi'(0)/E\left[\psi^2(\cdot)\right]\right) - 1.$$

For Huber's score function $\psi_c(r) = sign(r)\cdot\min(|r|, c)$, evaluated at the normal distribution we find that $\kappa = \tau_n^{-1} - 1 > 0$, for

$$\tau_n = E_\Phi\left[\psi_c^2(\cdot)\right] = 1 - 2\Phi(-c) - 2c\left(\phi(c) - c\Phi(-c)\right).$$

3. ROBUST DESIGN

In this section we obtain the maximum loss of the BLUP, for $f \in \mathcal{F}_\alpha, g \in \mathcal{G}_\beta, h \in \mathcal{H}_\gamma$, and discuss numerical approximations to be used in the repeated evaluation of this maximum, leading to minimax designs.

We propose to use those designs, optimized for use with the GLSE, even when $\hat{\boldsymbol{\theta}}$ is a GM-estimate. One reason for this is of course the relative intractability of the MSE of the GM-estimate, making it very difficult to maximize. A more compelling reason is that our experience has been that the robust designs seem to depend very little on the choice of the estimate - see for example Sinha and Wiens (2002).

Theorem 2. *For the BLUP of Theorem 1 the maximum value of* $\mathcal{L}(\mathbf{A}_0; f, g, h)$*, for* $f \in \mathcal{F}_\alpha, g \in \mathcal{G}_\beta, h \in \mathcal{H}_\gamma$ *is*

$$\max_{\mathcal{F}_\alpha, \mathcal{G}_\beta, \mathcal{H}_\gamma} \mathcal{L}(\mathbf{A}_0; f, g, h) =$$

$$\mathcal{L}(\mathbf{A}_0; f_0, g_0, h_0) + \frac{\sqrt{N}}{M}\left[\alpha \|\mathbf{d}_0\| + \beta\left\{tr\left(\left[\mathbf{B}_0\mathbf{B}_0^T\right]^2\right)\right\}^{1/2} + \gamma\lambda_{\max}\left(\mathbf{B}_0\mathbf{B}_0^T\right)\right], \quad (16)$$

where $\lambda_{\max}$ *denotes the largest eigenvalue. The maximum is attained if* $f(\mathbf{t}) = f_0(\mathbf{t})$ *for* $\mathbf{t} \notin \mathcal{S}$*,*

$$\mathbf{f}_1 = \mathbf{f}_{1,0} + \alpha\sqrt{N}\mathbf{d}_0/\|\mathbf{d}_0\|, \quad (17)$$

$$\mathbf{G} = \mathbf{G}_0 + \frac{\beta\sqrt{N}}{\left\{tr\left(\left[\mathbf{B}_0\mathbf{B}_0^T\right]^2\right)\right\}^{1/2}}\mathbf{B}_0^T\mathbf{B}_0, \quad (18)$$

and $\mathbf{h}$ *is an eigenvector of* $\mathbf{B}_0^T\mathbf{B}_0$*, with* $\|\mathbf{h}\|^2 = \gamma\sqrt{N}$*, corresponding to* $\lambda_{\max}$*.*

Note that if $M = 1$, as when X_{Total} is being predicted, then $\mathbf{B}_0$ is a row vector and in (16),

$$\left\{tr\left(\left[\mathbf{B}_0\mathbf{B}_0^T\right]^2\right)\right\}^{1/2} = \lambda_{\max}\left(\mathbf{B}_0\mathbf{B}_0^T\right) = \mathbf{B}_0\mathbf{B}_0^T.$$

3.1. Modifications for large M

The algorithms described in the next section call for the repeated calculation of the loss (16), hence of the eigenvalues of the $M \times M$ matrix $\mathbf{B}_0\mathbf{B}_0^T$. For large values of M this is not feasible in a reasonable amount of time. This in particular is a problem when the loss is (6) or (7), so that M is $N-n$ or N respectively, if N is realistically large. We have noticed however that in these cases $\lambda_{\max}\left(\mathbf{B}_0\mathbf{B}_0^T\right)$ is typically very close to the largest eigenvalue of the $n \times n$ matrix $\mathbf{P}_0^T\mathbf{P}_0$.

To explain this closeness, we first define $\mathbf{Q}_2$ to be the incidence matrix for $\mathcal{T}\backslash\mathcal{S}$, so that $\mathbf{Q} = \left(\mathbf{Q}_1^T : \mathbf{Q}_2^T\right)^T$ is an orthogonal matrix. For loss (6), $\mathbf{C} = \mathbf{Q}_2$ and so

$$\mathbf{Q}\mathbf{C}^T\mathbf{C}\mathbf{Q}^T = \mathbf{0}_{n\times n} \oplus \mathbf{I}_{N-n}. \quad (19)$$

For loss (7), $\mathbf{C} = \mathbf{I}_N$ and

$$\mathbf{Q}\mathbf{C}^T\mathbf{C}\mathbf{Q}^T = \mathbf{I}_N. \quad (20)$$

Lemma 1 shows that our approximation of the largest eigenvalue becomes exact as $\|\mathbf{F}_0\| \to 0, \infty$.

Lemma 1. *Assume that one of (19), (20) holds. Then*

$$\lambda_{\max}\left(\mathbf{B}_0\mathbf{B}_0^T\right) = \lambda_{\max}\left(\mathbf{P}_0^T\mathbf{P}_0\right) + o\left(1\right)$$

as either:
A1) $\|\mathbf{F}_0\| \to \infty$ *in such a way that* $\mathbf{\Sigma}_{11,0}^{-1} \to 0$ *and* $\mathbf{R}_0 \to \left(\mathbf{Z}_1^T\mathbf{Z}_1\right)^{-1}\mathbf{Z}_1^T$*, or*
A2) $\|\mathbf{F}_0\| \to 0$.

Conditions A1) and A2) hold if $\mathbf{F}_0 = \sigma^2\mathbf{I}_N$ and $\sigma^2 \to \infty$ and 0 respectively.

When (19) or (20) hold and $M > 25$, we replace $\lambda_{\max}\left(\mathbf{B}_0\mathbf{B}_0^T\right)$ by the much more easily computed $\lambda_{\max}\left(\mathbf{P}_0^T\mathbf{P}_0\right)$ in (16). This approximation is surprisingly accurate. As examples, we computed the relative error

$$re = \left|1 - \frac{\lambda_{\max}\left(\mathbf{P}_0^T\mathbf{P}_0\right)}{\lambda_{\max}\left(\mathbf{B}_0\mathbf{B}_0^T\right)}\right|$$

for various choices of N, and for a number of randomly chosen n-element subsets of $\mathcal{T}$. For $\mathbf{C} = \mathbf{I}_N$ and $(N, n) = (49, 5)$, $(100, 10)$ and $(400, 20)$, the average values of re over 100 trials were .05%, .12% and .05% respectively, when the regressors were $\mathbf{z}\left(\mathbf{t}\right) = (1, t_1, t_2)^T$ and $\mathbf{F}_0$, $\mathbf{G}_0$ were as described in §4.1 below. Varying the regression function or the choices of $\mathbf{F}_0$ and $\mathbf{G}_0$ typically resulted in even smaller relative errors.

A further simplification in evaluating the AMSPE, if the loss is (6) or (7), is to write

$$tr\left(\mathbf{B}_0\mathbf{G}_0\mathbf{B}_0^T\right) = \begin{cases} tr\left(\left[\mathbf{G}_{1,0}\mathbf{Q}_1^T\right]\left[\mathbf{A}_0^T\mathbf{A}_0 + 2\mathbf{Q}_1\mathbf{P}_0 - \mathbf{I}_n\right] - 2\mathbf{G}_{1,0}\mathbf{P}_0\right) + tr\left(\mathbf{G}_0\right), & \text{if } \mathbf{C} = \mathbf{Q}_2, \\ tr\left(\left[\mathbf{G}_{1,0}\mathbf{Q}_1^T\right]\left[\mathbf{A}_0^T\mathbf{A}_0\right] - 2\mathbf{G}_{1,0}\mathbf{A}_0\right) + tr\left(\mathbf{G}_0\right), & \text{if } \mathbf{C} = \mathbf{I}_N; \end{cases}$$

$$tr\left(\left[\mathbf{B}_0\mathbf{B}_0^T\right]^2\right) = \begin{cases} tr\left(\left[\mathbf{A}_0^T\mathbf{A}_0\right]^2 + 2\mathbf{A}_0^T\mathbf{A}_0\right) + N - n, & \text{if } \mathbf{C} = \mathbf{Q}_2, \\ tr\left(\left[\mathbf{A}_0^T\mathbf{A}_0\right]^2 + 4\mathbf{A}_0^T\mathbf{A}_0 + 2\left[\mathbf{Q}_1\mathbf{A}_0\right]^2\right. & \\ \left. -4\mathbf{Q}_1\mathbf{A}_0 - 4\mathbf{A}_0^T\mathbf{A}_0\mathbf{Q}_1\mathbf{A}_0\right) + N, & \text{if } \mathbf{C} = \mathbf{I}_N. \end{cases}$$

With the exception of $tr\left(\mathbf{G}_0\right)$, which must only be calculated once, the traces on the right hand sides are of $n \times n$ matrices.

4. DESIGNS: ALGORITHMS AND EXAMPLES

4.1. Test cases

We begin by exhibiting some optimal designs in situations in which N and n are small enough that the optimization can be carried out by an exhaustive search of all $\binom{N}{n}$ possible

designs. We consider two types of correlation structures. The first - Gaussian correlations - employs nominal covariances $g_0(\mathbf{t}, \mathbf{t}') = \sigma_2^2 \exp\left\{-\lambda \|\mathbf{t} - \mathbf{t}'\|^2\right\}$. The second - exponential correlations - uses $g_0(\mathbf{t}, \mathbf{t}') = \sigma_2^2 \exp\{-\lambda \|\mathbf{t} - \mathbf{t}'\|\}$. In our examples we set $\sigma_2 = 1$, and choose λ so that the nearest neighbour correlation is .8. The ideal error variances are taken to be $\sigma_1^2 \equiv 1$.

Figure 1 exhibits optimal designs with $N = 25$ and $n = 6$. In all cases we take $\mathbf{C} = \mathbf{I}_N$, so that we aim for estimation of AMSPE as at (7). Each of the designs in Figures 1(b) and 1(c) took about 180 seconds to compute (using MATLAB, on a 2200 MHz PC with 1 gigabyte of RAM). The modifications of §3.1 led to the same designs, in about 100 seconds. For Figure 1(a), in which the fitted model is in fact the correct one, these times were approximately halved. In all three cases, the same designs were obtained when the loss was given by (6).

4.2. Simulated annealing

We have found that simulated annealing can be quite successful in determining the optimal robust designs. Our algorithm is a modification of that of Sacks and Schiller (1988). It depends on a sequence $\{\pi_j\}$ of acceptance probabilities and parameters $\{n_0, \delta_0, \delta_1, \nu, m\}$, and is described as follows.

Suppose that at the j^{th} stage ($j = 0, 1, 2, ...$) of the process we are considering a configuration $S^{(j)} = \left\{\mathbf{t}_{i_1}^{(j)}, \ldots, \mathbf{t}_{i_n}^{(j)}\right\}$, with loss $L^{(j)}$ as at (16). Pick, at random, a location $\mathbf{t} \in \mathcal{T} \backslash S^{(j)}$ and determine in sequence the loss that arises if one of the $\mathbf{t}_{i_k}^{(j)}$ in $S^{(j)}$ is replaced by $\mathbf{t}$. If the least of these is less than $L^{(j)}$, then the corresponding configuration is accepted. Otherwise, this configuration is accepted with probability π_j.

Note that we are specifying that the same location not appear more than once in the design. Many practical situations, e.g. geological sampling, require this. It is not a crucial point however and the method works equally well when replicates are allowed.

For large values of N we sometimes modify this step by employing a suggestion of Royle (2002). The suggestion was made in connection with exchange algorithms but is also applicable to the current situation. In this "$100p$% nearest neighbour" modification we test only those $\mathbf{t}_{i_k}^{(j)}$ for which $\left\|\mathbf{t}_{i_k}^{(j)} - \mathbf{t}\right\| < p \max_{\mathbf{t}' \in \mathcal{T}} \|\mathbf{t}' - \mathbf{t}\|$.

The preceding step is repeated n_0 times, or until a new configuration is accepted, whichever comes first. If a new configuration is accepted it is labelled $S^{(j+1)}$, and its loss is labelled

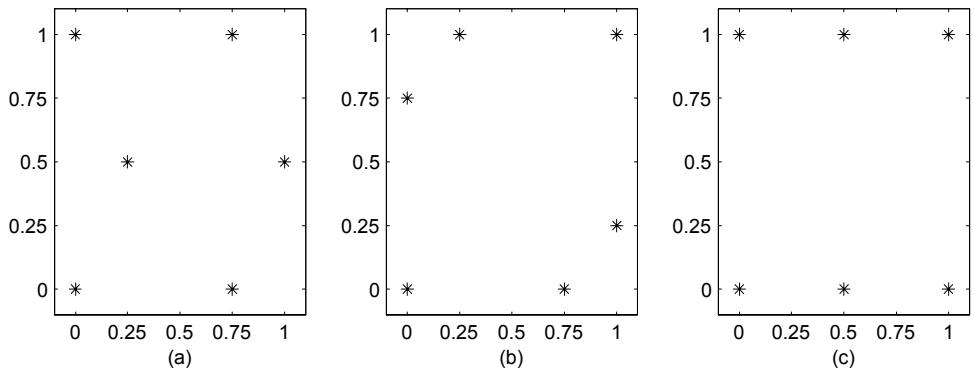

Figure 1: (a) Optimal scenario on an $N = 25$ point square grid in the unit square with $n = 6$ sites chosen. Inputs are $f_0(\mathbf{t}) \equiv \sigma_1^2 = 1$, Gaussian correlations $g_0(\mathbf{t}, \mathbf{t}') = \sigma_2^2 \exp\left\{-\lambda \|\mathbf{t} - \mathbf{t}'\|^2\right\}$ with $\sigma_2^2 = 1$, $\lambda = 3.5703$ for a nearest neighbour correlation of $e^{-.0625\lambda} = .8$. Regressors are $\mathbf{z}(\mathbf{t}) = (1, t_1, t_2)^T$ and $\mathbf{C} = \mathbf{I}_N$. Fitted model is correct: $\alpha = \beta = \gamma = 0$. The optimal sites are $\{1, 4, 12, 15, 21, 24\}$ respectively - counting left to right across row 1, then row 2, etc. Equivalent designs obtained by rotating this design through 90°, 180° and 270° have sites $\{1, 5, 8, 16, 20, 23\}$, $\{2, 5, 11, 14, 22, 25\}$ and $\{3, 6, 10, 18, 21, 25\}$ respectively. All result in a loss of 0.73358.
(b) As in (a), but now $\alpha = .25$, $\beta = \gamma = 1$. The optimal design shown has sites $\{1, 4, 10, 16, 22, 25\}$ and loss 3.2377, and is invariant under a rotation through 180°. The only other optimal design, obtained by rotating the one shown through 90° or 270°, has sites $\{2, 5, 6, 20, 21, 24\}$.
(c) As in (b), but with exponential correlations: $g_0(\mathbf{t}, \mathbf{t}') = \exp\{-\lambda \|\mathbf{t} - \mathbf{t}'\|\}$ with $\lambda = .8926$ for a nearest neighbour correlation of .8. The optimal design shown has sites $\{1, 3, 5, 21, 23, 25\}$ and loss 3.0897, and is invariant under a rotation through 180°. The only other optimal design, obtained by rotating the one shown through 90° or 270°, has sites $\{1, 5, 11, 15, 21, 25\}$.

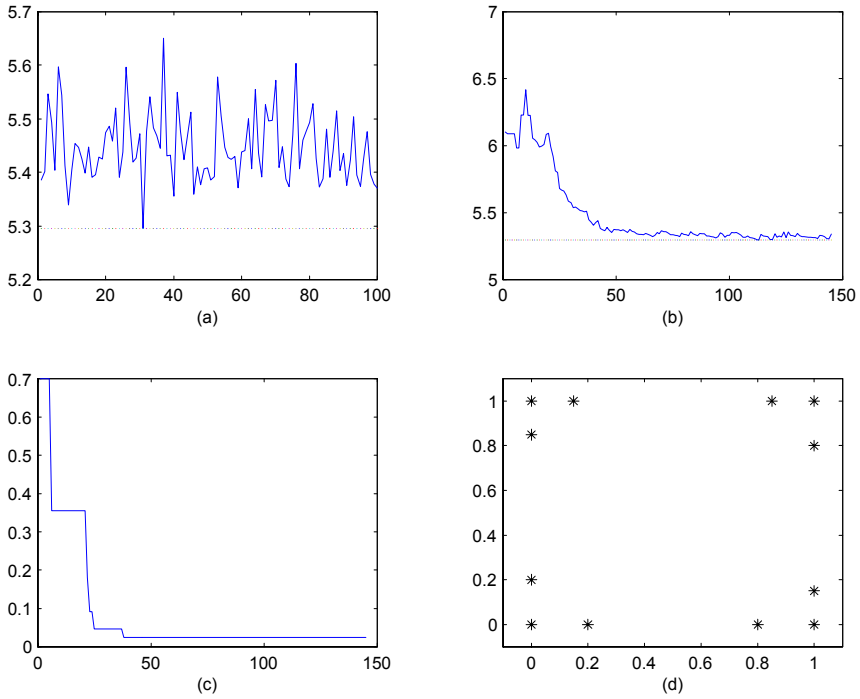

Figure 2: Output from 100 runs of the simulated annealing algorithm with the 20% nearest neighbour modification. Design space $\mathcal{T}$ is a 21×21 point grid from which $n = 12$ sites are chosen. Regressors are $\mathbf{z}(\mathbf{t}) = (1, t_1, t_2)^T$ and the parameters are $\alpha = .025$, $\beta = \gamma = 1$. (a) Minimum loss (AMSPE) vs. run. (b) Accepted loss vs. stage in best run; minimum loss = 5.2955. (c) Acceptance probabilities vs. stage in best run. (d) Best design found has sites {1, 5, 17, 21, 84, 85, 357, 358, 421, 424, 438, 441}.

$L^{(j+1)}$. If no new configuration is accepted, then $S^{(j)}$ is relabelled as $S^{(j+1)}$, $L^{(j)}$ as $L^{(j+1)}$. One then moves on to the next stage.

The sequence of acceptance probabilities is defined by $\pi_0 = .7$ and

$$\pi_{j+1} = \begin{cases} \min\left(1, \frac{\pi_j}{1-\delta_0}\right), & \text{if no new configuration was accepted at the } j^{th} \text{ stage,} \\ (1-\delta_1)\,\pi_j, & \text{if a new configuration was accepted,} \\ & \text{and } L^{(j+1)} < (1-\nu)\min_{i \le j} L^{(i)}, \\ \pi_j, & \text{otherwise.} \end{cases}$$

Iterations cease when there have been m evaluations of the loss since the last change in the value of the acceptance probability. The initial state $S^{(0)}$ is the best of m randomly chosen configurations.

This describes one "run" of the annealing algorithm. We have found it best to carry

out a number of runs, with different parameter values. Typically, in a run we will randomly choose $n_0 \in [.1\,(N-n)\,, .5\,(N-n)]$, $\delta_0 \in [.1, .5]$, $\delta_1 \in [.3, .5]$, $\nu \in [.01, .05]$ and $m \in [50, 200]$. For the situations illustrated in Figure 1, the optimal configuration was generally found in no more than 4 runs. To carry out four runs requires about 15 seconds of computing time.

Figure 2 illustrates the output from a set of 100 runs of the annealing algorithm, with $N = 441$ points arranged in a square grid, from which $n = 12$ locations are chosen. There are $\binom{441}{12} \approx 10^{23}$ possible designs. The parameters used were $\alpha = .025$, $\beta = \gamma = 1$, the correlations were of the Gaussian type, loss was AMSPE and the regressors were $\mathbf{z}(\mathbf{t}) = (1, t_1, t_2)^T$. The 20% nearest neighbour modification to the annealing algorithm was employed, as were the simplifications of §3.1. Each run required 5.6 seconds of computing time and 342 evaluations of the loss, on average.

Although it is time consuming, we can successfully run the annealing algorithm with inputs at least as large as $n = 100$, $N = 5000$. Of course designs for predicting X_{Total} (see Example 1.1) can be obtained much more quickly.

4.3. Sequential design

Choosing the design points sequentially is an obvious alternative, and one which we have also investigated. One "run" of our algorithm consists of randomly choosing p points from $\mathcal{T}$, then finding that $(p+1)^{th}$ point which minimizes the loss when appended to the current p-point design, and repeating until n points have been determined. We run the algorithm numerous times, and choose the best of the resulting designs.

For each of the scenarios of Figure 1, we ran the sequential procedure 190 times, thus using about the same amount of time as 4 runs of the annealing algorithm. Although the optimal designs were rarely found in this way, the sequentially determined designs were at least close to optimal. The amounts by which their loss exceeded the minimum loss, expressed as a percentage of the minimum loss, were typically $< 2\%$.

For large values of N it does not seem feasible to carry out many runs of the sequential approach. There is no substitute here for the nearest neighbour modification, which drastically reduces the number of evaluations of the loss which are required in the annealing algorithm. Figure 3 shows the result of 15 runs of the sequential procedure using the same inputs as in Figure 2. Each run required 3907 evaluations of the loss, and the total sequence

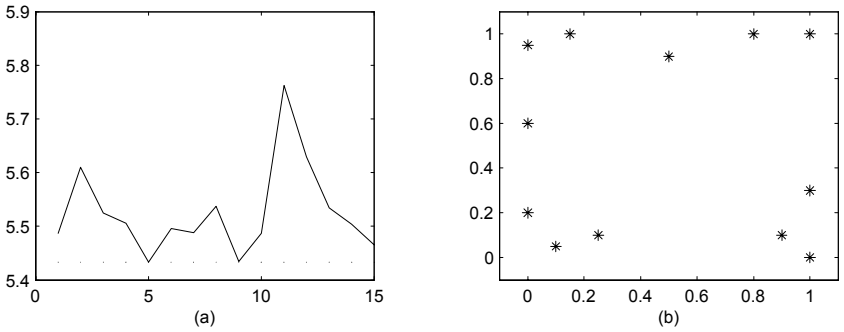

Figure 3: Output from 15 runs of the sequential procedure. Inputs are the same as for Figure 2. (a) Loss vs. run. (b) Best design found, with loss 5.4327, has sites {21, 24, 48, 61, 85, 147, 253, 389, 400, 424, 437,441}.

of runs required 10 minutes of computing time - the same as 100 runs of the simulated annealing algorithm. The loss for the best design found was 5.4327, exceeding that in Figure 2 by 2.6%.

5. ROBUST ESTIMATION AND PREDICTION: DETAILS AND SIMULATIONS

The preceding development has assumed a known covariance structure. In practice, one would typically posit a parametric form for this structure, and substitute parameter estimates. These might be obtained from a preliminary or previous study, and could then be used prior to constructing a static design for the current study. Alternatively the estimation and design steps could be carried out iteratively and sequentially. In any event, robust estimation methods are required.

Robust methods for variogram estimation were studied by Cressie and Hawkins (1980) and by Genton (2001; see also references to earlier work therein). Both employ judiciously chosen order statistics of the differences $|X(\mathbf{t}) - X(\mathbf{t}')|$, or the second differences. (In particular, it is assumed in the references of this paragraph that $X(\mathbf{t})$ is observable, i.e. that $\sigma_1^2 = 0$.) Nonparametric covariogram estimation is studied by Genton and Gorsich (2002). Militino and Ugarte (1997) propose 1-step Schweppe-type GM-estimation of regression, after applying a transformation, based on the residuals from an initial least trimmed squares fit which ignores the dependence structure, to achieve an approximately diagonal covariance matrix.

We propose here a method of GM-estimation of the regression parameters, and correspondingly a robust method of prediction, appropriate when measurement errors are present. We suppose that the experimenter assumes the measurement errors $\varepsilon(\mathbf{t})$ to have common variance σ_1^2, and the covariance function to be of the form

$$g(\mathbf{t}, \mathbf{t}') = \sigma_2^2 \rho_\lambda \left(\|\mathbf{t} - \mathbf{t}'\| \right),$$

for an isotropic correlation function ρ_λ depending on a, possibly multidimensional, parameter λ. Under the assumed regression model, the data vector $\mathbf{y}$ has mean $\mathbf{Z}_1\boldsymbol{\theta}$ and covariance matrix $\sigma_1^2\mathbf{S}$, where $\mathbf{Z}_1$ has rows $\mathbf{z}^T(\mathbf{t}_{i_1}), \ldots, \mathbf{z}^T(\mathbf{t}_{i_n})$ and $\mathbf{S} = \mathbf{I}_n + \zeta\boldsymbol{\Phi}$ for $\zeta = \sigma_2^2/\sigma_1^2$ and $\boldsymbol{\Phi}_{jk} = \rho_\lambda\left(\left\|\mathbf{t}_{i_j} - \mathbf{t}_{i_k}\right\|\right)$.

Regression/scale step ("R/S"): Given a trial value $\hat{\mathbf{S}} = \mathbf{I}_n + \hat{\zeta}\hat{\boldsymbol{\Phi}}$, put $\mathbf{v} = \hat{\mathbf{S}}^{-1/2}\mathbf{y}$, $\mathbf{U} = \hat{\mathbf{S}}^{-1/2}\mathbf{Z}_1$ and obtain trial estimates $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{GM}$ and $\hat{\sigma}_1 = \hat{\sigma}_n$ as at (12) and (13), with $\psi_i(r) = w(\mathbf{u}_i)\,\psi_{c_1}(r)$, $\chi(r) = \psi_{c_1}^2(r)$ and τ_n as in Remark 6, with $c = c_1$. In our simulations the only nonconstant regressors are the locations $\mathbf{t}$, so that influential carriers are not an issue. Thus we take constant weights $w(\mathbf{u}_i) \equiv 1$. We "solve" (12) and (13) by carrying out three iterations of reweighted least squares, in each step.

Covariance step ("C"): Given a trial value $\hat{\boldsymbol{\theta}}$, set $\mathbf{r} = \left(\mathbf{y} - \mathbf{Z}_1\hat{\boldsymbol{\theta}}\right)/\hat{\sigma}_1$ and let the robustified residual vector $\boldsymbol{\psi}$ have elements $\psi_{c_2}(r_i)$. Minimize, over ς and λ, the function

$$\log(\det(\mathbf{S})) + \xi_{c_2}\left(\left\|\mathbf{S}^{-1/2}\mathbf{r}\right\|\right),$$

where

$$\xi_c(t) = 2\int_0^t \psi_c(x)dx = \begin{cases} t^2, & |t| \le c, \\ 2c|t| - c^2, & |t| \ge c. \end{cases}$$

If the likelihood is Gaussian, then these steps yield the maximum likelihood estimates when $c_1 = c_2 = \infty$. The covariance step gives the MLE of ς and λ if the density of $\mathbf{r}$ is of the form $|\mathbf{S}|^{-1/2}\,p\left(\left\|\mathbf{S}^{-1/2}\mathbf{r}\right\|\right)$ for $p(t) \propto \exp\left(-\xi_c(t)/2\right)$. We have, at least initially, used $c_1 = 1.5, c_2 = \sqrt{n}$. In order to keep the estimate $\hat{\varsigma}$ away from zero, we have sometimes found it effective to increase these values somewhat - see the example in §6.

The MATLAB function *fmincon* carries out the minimization in step C quickly and effectively. It is generally sufficient to iterate five times between R/S and C.

Table 1. Simulation results for M-estimation and prediction. Regression biases and prediction errors, with standard errors in parentheses. True and assumed correlations both of the exponential form.

Errors/ Regression[1]	Method[2]	$\%RNB$		$\%ARPE$		$ASPE$	
		Nearest neighbour correlation = .2					
U/U	M	12.49	(.81)	4.34	(.13)	1.07	(.07)
	GLS	11.96	(.77)	4.24	(.12)	1.03	(.06)
	M+D	7.89	(.36)	3.73	(.06)	.79	(.03)
	GLS+D	7.79	(.35)	3.72	(.06)	.78	(.03)
U/C	M	13.35	(.80)	4.78	(.13)	1.36	(.09)
	GLS	13.03	(.73)	4.69	(.12)	1.29	(.07)
	M+D	8.14	(.37)	4.07	(.08)	.93	(.04)
	GLS+D	7.95	(.39)	3.99	(.08)	.90	(.03)
C/U	M	13.73	(.72)	4.67	(.13)	1.30	(.08)
	GLS	16.83	(1.82)	5.23	(.36)	2.06	(.59)
	M+D	8.49	(.40)	3.84	(.07)	.84	(.03)
	GLS+D	9.87	(.69)	4.19	(.16)	1.09	(.13)
C/C	M	15.89	(1.72)	5.24	(.27)	2.02	(.48)
	GLS	23.68	(3.24)	6.82	(.64)	4.43	(1.18)
	M+D	8.64	(.50)	4.15	(.09)	1.01	(.05)
	GLS+D	13.21	(1.56)	5.38	(.44)	2.77	(.79)
		Nearest neighbour correlation = .8					
U/U	M	12.60	(.71)	3.87	(.12)	.91	(.08)
	GLS	12.37	(.68)	3.79	(.11)	.87	(.08)
	M+D	8.80	(.44)	3.59	(.10)	.72	(.04)
	GLS+D	8.89	(.44)	3.55	(.09)	.70	(.04)
U/C	M	12.87	(.68)	4.16	(.12)	1.04	(.06)
	GLS	12.36	(.67)	4.07	(.12)	1.00	(.06)
	M+D	9.27	(.49)	3.73	(.09)	.79	(.04)
	GLS+D	8.87	(.46)	3.62	(.09)	.74	(.03)
C/U	M	14.66	(.95)	4.36	(.15)	1.27	(.21)
	GLS	23.63	(5.53)	6.08	(.95)	8.92	(5.31)
	M+D	8.44	(.44)	3.65	(.12)	.77	(.06)
	GLS+D	14.56	(2.95)	5.31	(.83)	4.66	(2.54)
C/C	M	14.84	(.85)	4.39	(.16)	1.20	(.10)
	GLS	26.15	(6.38)	7.55	(1.75)	19.07	(14.21)
	M+D	10.17	(.44)	3.87	(.09)	.86	(.04)
	GLS+D	19.22	(4.98)	6.61	(1.56)	14.91	(11.23)

1. U = Uncontaminated, C = Contaminated. 2. "+D" = sites chosen by design.

Table 2. Simulation results for M-estimation and prediction. Regression biases and prediction errors, with standard errors in parentheses. True correlations exponential, assumed correlations Gaussian.

Errors/ Regression[1]	Method[2]	$\%RNB$		$\%ARPE$		$ASPE$	
		Nearest neighbour correlation = .2					
U/U	M	12.25	(.67)	4.54	(.12)	1.18	(.06)
	GLS	12.04	(.63)	4.49	(.11)	1.15	(.06)
	M+D	7.94	(.37)	3.91	(.09)	.87	(.04)
	GLS+D	7.78	(.37)	3.89	(.08)	.85	(.03)
U/C	M	12.32	(.65)	4.53	(.11)	1.20	(.06)
	GLS	12.25	(.67)	4.50	(.11)	1.18	(.06)
	M+D	8.13	(.37)	3.97	(.07)	.89	(.03)
	GLS+D	8.10	(.41)	3.94	(.07)	.88	(.03)
C/U	M	12.87	(.76)	4.83	(.33)	1.72	(.46)
	GLS	25.53	(6.29)	7.89	(1.65)	14.45	(9.91)
	M+D	8.14	(.41)	3.80	(.07)	.84	(.03)
	GLS+D	18.06	(4.85)	6.46	(1.32)	13.76	(10.51)
C/C	M	13.59	(.83)	4.85	(.17)	1.62	(.32)
	GLS	18.94	(2.34)	5.96	(.51)	3.54	(1.15)
	M+D	8.41	(.41)	4.09	(.08)	.96	(.04)
	GLS+D	11.71	(1.26)	5.08	(.38)	2.21	(.61)
		Nearest neighbour correlation = .8					
U/U	M	13.98	(.81)	4.07	(.14)	1.01	(.09)
	GLS	13.64	(.82)	4.02	(.13)	.97	(.08)
	M+D	8.26	(.37)	3.61	(.09)	.74	(.03)
	GLS+D	8.16	(.36)	3.51	(.08)	.71	(.03)
U/C	M	13.75	(.85)	4.28	(.14)	1.11	(.07)
	GLS	13.58	(.85)	4.17	(.13)	1.06	(.07)
	M+D	9.52	(.48)	3.72	(.10)	.80	(.04)
	GLS+D	9.28	(.46)	3.67	(.10)	.78	(.04)
C/U	M	13.29	(.93)	4.46	(.25)	1.33	(.24)
	GLS	14.89	(1.54)	4.68	(.28)	1.60	(.36)
	M+D	8.92	(.43)	3.73	(.10)	.79	(.04)
	GLS+D	10.40	(.63)	4.07	(.17)	1.02	(.11)
C/C	M	15.66	(2.26)	4.82	(.55)	4.76	(3.64)
	GLS	27.68	(7.16)	7.32	(1.36)	14.73	(9.46)
	M+D	9.64	(.47)	3.83	(.11)	.86	(.06)
	GLS+D	16.28	(2.47)	5.71	(.69)	4.62	(1.76)

1. U = Uncontaminated, C = Contaminated. 2. "+D" = sites chosen by design.

A small simulation study was carried out to test these methods. The main conclusions remained constant across a range of inputs, and are reported in detail here for the following cases. We considered an $N = 10 \times 20$ grid of equally spaced locations. We simulated 100 populations of size N, and from each randomly chose a sample of size $n = 15$. The distribution of the measurement error ε was either $N(0, \sigma_1^2 = 2)$ ("uncontaminated errors") or ("contaminated errors") was this distribution mixed with 1 "slash" error - a $N(0, \sigma_1^2 = 2)$ variable divided by an independent uniform $U[0, 1]$ variable - per sample. The marginal distribution of $\delta(\mathbf{t})$ was $N(0, \sigma_2^2 = .5)$. Thus $\zeta = .25$. The true correlations were of an exponential form, with $\rho_\lambda(d) = e^{-\lambda d}$ and λ chosen for a nearest neighbour correlation of either .2 or .8. The fitted regression model used regressors $\mathbf{z}(\mathbf{t}) = (1, t_1, t_2)^T$. The true regression response was either $E[X(\mathbf{t})] = \mathbf{z}^T(\mathbf{t})\boldsymbol{\theta}$, with $\boldsymbol{\theta} = (10, 10, 10)^T$, ("uncontaminated regression") or was $E[X(\mathbf{t})] = \mathbf{z}^T(\mathbf{t})\boldsymbol{\theta} + h(\mathbf{t})$ ("contaminated regression"), with $h(\mathbf{t}) \propto t_1 t_2 + a t_1 + b t_2 + c$ and the constants chosen to satisfy (2) and $\sum_{\mathbf{t}\in\mathcal{T}} h^2(\mathbf{t}) = \sqrt{N}$. Note that (2) ensures that $\boldsymbol{\theta}$, as above, is still the best fitting parameter, in the sense of (1), even under a contaminated regression. With contaminated errors or covariances however the "best" parameters $(\sigma_1^2, \sigma_2^2, \lambda)$ are no longer those used in the simulations. Thus we base our comparisons on the accuracy in the estimation of $\boldsymbol{\theta}$, and on the accuracy of the predictions of $X(\mathbf{t})$. The former is gauged by the percent relative norm of the bias of $\hat{\boldsymbol{\theta}}$:

$$\%RNB = 100 \cdot \frac{\left\| \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right\|}{\|\boldsymbol{\theta}\|},$$

together with the associated standard errors. The latter is gauged by the percent average relative prediction error:

$$\%ARPE = 100 \cdot aver_{\mathbf{t}\in\mathcal{T}} \left(\left| \frac{\hat{X}(\mathbf{t}) - X(\mathbf{t})}{X(\mathbf{t})} \right| \right),$$

with $\hat{X}(\mathbf{t})$ computed as at (15) with $\mathbf{C} = \mathbf{I}_N$, and by the average squared prediction error:

$$ASPE = aver_{\mathbf{t}\in\mathcal{T}} \left(\left(\hat{X}(\mathbf{t}) - X(\mathbf{t}) \right)^2 \right).$$

From each sample the M-estimate and associated predictions were computed. This was then repeated, using the same sample but with the GLS estimate. Both sets of estimates were then recomputed, using the same populations but now using designs determined as in

§4. In constructing these designs we took $\alpha = \beta = \gamma = 1$, and used the true values of σ_1^2, σ_2^2 and of the nearest neighbour correlation. The assumed correlation structure used was the same as in step C of the estimation procedure. The designs are quite stable under changes in the correlation structure - see Figure 4.

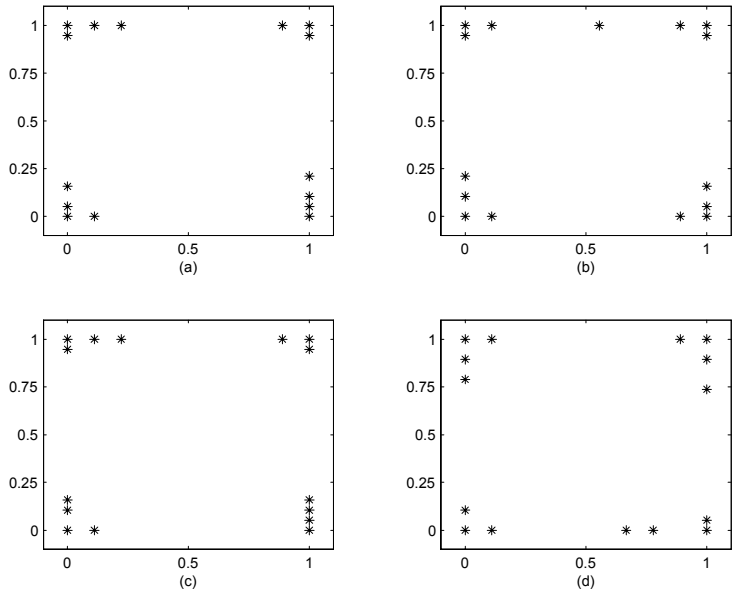

Figure 4: Designs for simulation study of §5. Correlation structures and nearest neighbour correlations are (a) Exponential, .2; (b) Exponential, .8; (c) Gaussian, .2; (b) Gaussian, .8.

The results from the simulations are presented in Table 1, for the case in which the true and assumed correlation structures coincided, and in Table 2 for the case in which the assumed structure was Gaussian - $\rho_\lambda(d) = e^{-\lambda d^2}$. In both cases the benefits of the robust estimation were considerable when the errors were contaminated. When only the regression was contaminated the GLSE was slightly superior, with the difference largely disappearing when the sites were chosen by design, rather than at random. For both estimation procedures, the benefits of the design were stronger when the assumed and true correlations structures differed.

The simulations were also run with 2 slash errors per sample (not shown). This often resulted in almost complete breakdown of the GLSE, while the M-estimates remained stable.

Table 3. Parameter estimates (standard errors in parentheses) for coal-ash study.

n	$\hat{\theta}_0$		$\hat{\theta}_1$		$\hat{\theta}_2$		$\hat{\sigma}_1^2$	$\hat{\varsigma}$	$\hat{\lambda}$
				M-estimates					
10	10.99	(.180)	−.19	(.017)	.06	(.013)	.19	.39	1.00
30	10.59	(.059)	−.13	(.006)	.026	(.004)	.42	.27	2.92
				GLS estimates					
10	10.99	(.181)	−.19	(.017)	.06	(.013)	.13	.95	1.37
30	10.60	(.063)	−.09	(.006)	.01	(.004)	.34	.77	14.68

6. EXAMPLE

We have restudied the "coal-ash" data, given by Gomez and Hazen (1970) and described in Cressie (1991). There are 208 coal-ash core measurements obtained from the Pittsburgh coal seam, at sites throughout a grid as displayed in Figure 5. The object of our study is to obtain an efficient and robust design upon which to base regression estimates of the effects in the east-west and north-south directions, corresponding to positive and negative values of t_1 and t_2 respectively. An initial ten-point design, given in Figure 5(a), was chosen. An approximate exponential correlation model was decided upon, and we then fitted a regression model, with regressors $(1, t_1, t_2)$, to these data. The initial M-estimates, with tuning constants $c_1 = 1.5, c_2 = \sqrt{n}$, had $\hat{\varsigma} = 0$. Upon increasing these to $c_1 = 2, c_2 = 2\sqrt{n}$ we obtained the values shown in Table 3. The corresponding standard errors of $\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2$, determined from the asymptotic covariance matrix

$$asym.cov\left[\hat{\boldsymbol{\theta}}\right] = \frac{\sigma_1^2}{n} \frac{E\left[\psi^2\left(\frac{\eta}{\sigma_1}\right)\right]}{\left(E\left[\psi'\left(\frac{\eta}{\sigma_1}\right)\right]\right)^2} \left(\mathbf{U}_{1,0}^T \mathbf{U}_{1,0}\right)^{-1}, \tag{21}$$

are also given in Table 3. The expectations in (21) were estimated by the corresponding sample averages (using "$n - p$" as the divisor for the expectation in the numerator). The GLS estimates and standard errors were very similar.

We then determined, by simulated annealing, robust designs ($\alpha = \beta = \gamma = 1$) consisting of a further 20 points chosen to minimize the maximum (estimated) loss (16). Despite the similarity in the estimates, the corresponding designs - see Figures 5(b), 5(c) - were slightly different. The parameters were then re-estimated. Note the implausibly large GLS

estimate $\hat{\lambda} = 14.68$, implying that even observations from nearest neighbours are essentially uncorrelated. This difference in the estimates perhaps accounts as well for the somewhat different predictions - see Figure 6, where the GLS predicted values, and those obtained from the M-estimates, are plotted against each other.

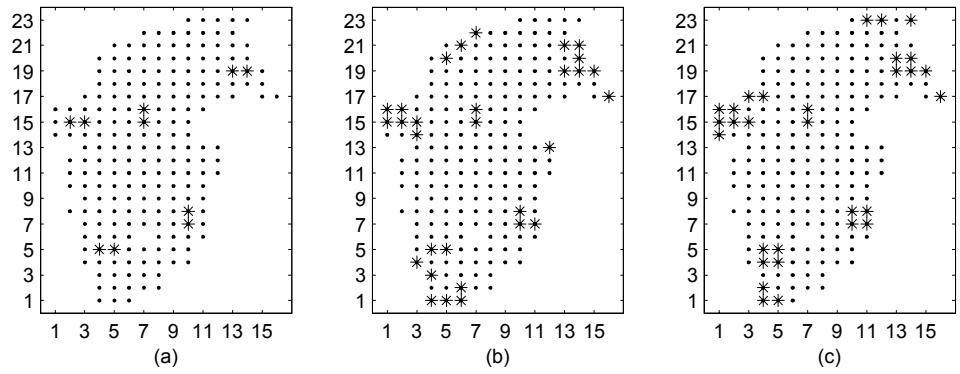

Figure 5: Designs for the example of §6: (a) initial design, (b) final design based on initial M-estimates, (c) final design based on initial GLS estimates.

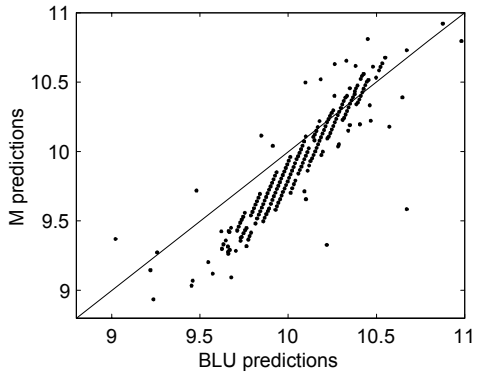

Figure 6: M-estimation based coal-ash predictions against the sorted BLU predictions.

7. SUMMARY

In this article we have derived robust methods for the design, estimation and prediction of spatial processes. Our framework assumes that the stochastic process of interest is itself subject to measurement error, and has a mean structure relying on a set of regressors. The measurement error variances, the correlations between observations made at differing

locations, and the regression structure are all typically only partially known, and may be incorrectly specified by the experimenter. We have exhibited the Best Linear Unbiased Predictor, and maximized a loss function, based on the mean squared error of this predictor, over neighbourhoods quantifying the various sources of model uncertainty. Two algorithms - one using simulated annealing and the other sequential in nature - have been introduced in order to minimize the maximum loss, leading to minimax designs. Parameter estimation and process prediction methods have also been introduced. These are based on generalized M-estimators, and are robust against contaminated error distributions.

A simulation study has shown that the procedures perform much as hoped, affording a substantial level of robustness when these model inadequacies are present, while being almost as efficient as more classical methods otherwise.

APPENDIX: DERIVATIONS

Proof of Theorem 1: The method is as in Cressie (1991). We find that

$$\mathbf{MSE}\left(\mathbf{A}; f, g, h\right) = \left(\mathbf{AQ}_1 - \mathbf{C}\right)\left[\left(\mathbf{Z}\boldsymbol{\theta} + \mathbf{h}\right)\left(\mathbf{Z}\boldsymbol{\theta} + \mathbf{h}\right)^T + \mathbf{G}\right]\left(\mathbf{AQ}_1 - \mathbf{C}\right)^T + \mathbf{AQ}_1\mathbf{FQ}_1^T\mathbf{A}^T, \tag{A.1}$$

so that

$$\mathbf{MSE}\left(\mathbf{A}; f_0, g_0, h_0\right) = \left(\mathbf{AQ}_1 - \mathbf{C}\right)\left[\left(\mathbf{Z}\boldsymbol{\theta}\right)\left(\mathbf{Z}\boldsymbol{\theta}\right)^T + \mathbf{G}_0\right]\left(\mathbf{AQ}_1 - \mathbf{C}\right)^T + \mathbf{AQ}_1\mathbf{F}_0\mathbf{Q}_1^T\mathbf{A}^T$$

and we seek

$$\mathbf{A}_0 = \arg\min_{\mathbf{A}} tr\left[\mathbf{MSE}\left(\mathbf{A}; f_0, g_0, h_0\right) - 2\left(\mathbf{AQ}_1 - \mathbf{C}\right)\mathbf{Z}\boldsymbol{\Lambda}^T\right],$$

where $\boldsymbol{\Lambda}$ is an $M \times p$ matrix of Lagrange multipliers.

The first order conditions are

$$\left(\mathbf{AQ}_1 - \mathbf{C}\right)\mathbf{Z}\boldsymbol{\theta}\boldsymbol{\theta}^T\mathbf{Z}^T\mathbf{Q}_1^T + \left(\mathbf{AQ}_1 - \mathbf{C}\right)\mathbf{G}_0^T\mathbf{Q}_1^T + \mathbf{AQ}_1\mathbf{F}_0^T\mathbf{Q}_1^T - \boldsymbol{\Lambda}\mathbf{Z}^T\mathbf{Q}_1^T = \mathbf{0}_{M\times n} \tag{A.2}$$

together with (8), by which the first term in (A.2) vanishes. This results in

$$\mathbf{A} = \left(\mathbf{CG}_{1,0}^T + \boldsymbol{\Lambda}\mathbf{Z}_1^T\right)\boldsymbol{\Sigma}_{11,0}^{-1}.$$

Substituting this into (8) gives

$$\boldsymbol{\Lambda} = \mathbf{C}\left(\mathbf{Z} - \mathbf{G}_{1,0}^T\boldsymbol{\Sigma}_{11,0}^{-1}\mathbf{Z}_1\right)\left(\mathbf{Z}_1\boldsymbol{\Sigma}_{11,0}^{-1}\mathbf{Z}_1\right)^{-1}$$

and (9) follows. Then from (A.1),

$$tr\left[\mathbf{MSE}\left(\mathbf{A}_0; f, g, h\right)\right] = tr\mathbf{A}_0^T\mathbf{A}_0\mathbf{F}_{11} + tr\mathbf{B}_0^T\mathbf{G}\mathbf{B}_0 + \left\|\mathbf{B}_0\mathbf{h}\right\|^2.$$

Since $\mathbf{F}_{11}$ is diagonal, we may replace $\mathbf{A}_0^T\mathbf{A}_0$ by its diagonal, obtaining (10). □

Derivation of (15): Define an $n \times 1$ vector $\mathbf{p} = \sigma_n\left(..., \psi_i\left(\eta_i/\sigma_n\right), ...\right)^T$, and let $\mathbf{E}_1$ – $\mathbf{E}_3$ be the $n \times n$ diagonal matrices with diagonal elements $E\left[\psi_i'\left(\eta_i/\sigma_n\right)\right]$, $E\left[\psi_i^2\left(\eta_i/\sigma_n\right)\right]$, $\psi_i'\left(0\right)$ respectively. Under appropriate conditions - see for example Wiens (1996) - there is an expansion of the form

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{GM} - \boldsymbol{\theta}\right) = \left(n^{-1}\mathbf{U}_{1,0}^T\mathbf{E}_1\mathbf{U}_{1,0}\right)^{-1}\left(n^{-1}\mathbf{U}_{1,0}^T\mathbf{E}_1\mathbf{k}_1 + n^{-1/2}\mathbf{U}_{1,0}^T\mathbf{p}\right) + o_P\left(1\right). \tag{A.3}$$

Upon replacing $\hat{\boldsymbol{\theta}}_{GLS}$ by $\hat{\boldsymbol{\theta}}_{GM}$ and $\mathbf{e}$ by $\mathbf{S}_{11,0}^{1/2}\hat{\mathbf{p}}$, equations (14) become

$$\mathbf{C}E\left[\left(\hat{\mathbf{x}} - \mathbf{x}\right)\hat{\mathbf{p}}^T\right] + \mathbf{V}\mathbf{S}_{11,0}^{1/2}E\left[\hat{\mathbf{p}}\hat{\mathbf{p}}^T\right] = \mathbf{0}_{M\times n}, \tag{A.4}$$

and our robustified predictor is

$$\left(\widehat{\mathbf{C}\mathbf{x}}\right)_{GM} = \mathbf{C}\mathbf{Z}\hat{\boldsymbol{\theta}}_{GM} + \mathbf{V}_{GM}\mathbf{S}_{11,0}^{1/2}\hat{\mathbf{p}}.$$

Taking $(f, g, h) = (f_0, g_0, 0)$ we approximate $E\left[\hat{\mathbf{p}}\hat{\mathbf{p}}^T\right]$ by $E_\Phi\left[\mathbf{p}_0\mathbf{p}_0^T\right] = \sigma_n^2\mathbf{E}_2$. Using (A.3) (with $\mathbf{k}_1 = \mathbf{0}$) and an approximation to $E\left[\mathbf{p}_0\boldsymbol{\delta}_0^T\right]$ given at (A.5) below, we approximate

$$E\left[\left(\hat{\mathbf{x}} - \mathbf{x}\right)\hat{\mathbf{p}}^T\right] = \mathbf{Z}E\left[\left(\hat{\boldsymbol{\theta}}_{GM} - \boldsymbol{\theta}\right)\hat{\mathbf{p}}^T\right] - \left(E\left[\mathbf{p}\boldsymbol{\delta}^T\right]\right)^T$$

by

$$\mathbf{Z}\left(\mathbf{U}_{1,0}^T\mathbf{E}_1\mathbf{U}_{1,0}\right)^{-1}\mathbf{U}_{1,0}^T E_\Phi\left[\mathbf{p}\mathbf{p}^T\right] - \mathbf{G}_{1,0}^T\mathbf{S}_{11,0}^{-1/2^T}\mathbf{E}_3.$$

With $\mathbf{K} = \mathbf{E}_3\mathbf{E}_2^{-1} - \mathbf{I}$ this gives

$$\mathbf{V}_{GM} = \sigma_n^{-2}\mathbf{C}\mathbf{G}_{1,0}^T\mathbf{S}_{11,0}^{-1} + \mathbf{C}\left[\sigma_n^{-2}\mathbf{G}_{1,0}^T\mathbf{S}_{11,0}^{-1/2^T}\mathbf{K} - \mathbf{Z}\left(\mathbf{U}_{1,0}^T\mathbf{E}_1\mathbf{U}_{1,0}\right)^{-1}\mathbf{U}_{1,0}^T\right]\mathbf{S}_{11,0}^{-1/2}$$

and

$$\left(\widehat{\mathbf{C}\mathbf{x}}\right)_{GM} = \mathbf{C}\left[\mathbf{Z}\hat{\boldsymbol{\theta}}_{GM} + \sigma_n^{-2}\mathbf{G}_{1,0}^T\mathbf{S}_{11,0}^{-1/2^T}\hat{\mathbf{p}}\right] + \mathbf{C}\left[\sigma_n^{-2}\mathbf{G}_{1,0}^T\mathbf{S}_{11,0}^{-1/2^T}\mathbf{K} - \mathbf{Z}\left(\mathbf{U}_{1,0}^T\mathbf{E}_1\mathbf{U}_{1,0}\right)^{-1}\mathbf{U}_{1,0}^T\right]\hat{\mathbf{p}}.$$

From (12), $\left(\mathbf{U}_{1,0}^T\mathbf{E}_1\mathbf{U}_{1,0}\right)^{-1}\mathbf{U}_{1,0}^T\hat{\mathbf{p}} = o_P(n^{-1/2})$, whence we ignore this term and obtain (15).

The approximation to $E\left[\mathbf{p_0}\boldsymbol{\delta}_0^T\right]$ arises as follows. Write the $(i,j)^{th}$ element as $\sigma_n E\left[\psi_i\left(\frac{\boldsymbol{\alpha}_i^T(\boldsymbol{\varepsilon}+\boldsymbol{\delta})}{\sigma_n}\right)\delta_j\right]$, where $\boldsymbol{\alpha}_i^T$ is the i^{th} row of $\mathbf{S}_{11,0}^{-1/2}\mathbf{Q}_1$. Expanding around $\boldsymbol{\varepsilon}_0+\boldsymbol{\delta}_0=\mathbf{0}$ and terminating the expansion after the linear term gives

$$E\left[\mathbf{p}_0\boldsymbol{\delta}_0^T\right]_{i,j} \approx \sigma_n E\left[\psi_i(0)\,\delta_j\right] + E\left[\psi_i'(0)\left(\boldsymbol{\alpha}_i^T\left(\boldsymbol{\varepsilon}_0+\boldsymbol{\delta}_0\right)\right)\delta_j\right].$$

and so

$$E\left[\mathbf{p}_0\boldsymbol{\delta}_0^T\right] \approx \mathbf{E}_3\mathbf{S}_{11,0}^{-1/2}\mathbf{G}_{1,0}. \tag{A.5}$$

□

Proof of Theorem 2: To carry out the maximizations over $\mathcal{F}_\alpha$, $\mathcal{G}_\beta$ and $\mathcal{H}_\gamma$ we shall first ignore the non-negativity constraints in the definitions of $\mathcal{F}_\alpha$ and $\mathcal{G}_\beta$, and the orthogonality constraint in the definition of $\mathcal{H}_\gamma$. We will then verify that the unconstrained maximizers also satisfy the constraints.

For $\mathcal{F}_\alpha$ we have, by the Cauchy-Schwarz inequality, that

$$\begin{aligned}\mathbf{d}_0^T\mathbf{f}_1 &= \mathbf{d}_0^T\mathbf{f}_{1,0} + \mathbf{d}_0^T\left(\mathbf{f}_1-\mathbf{f}_{1,0}\right)\\ &\leq \mathbf{d}_0^T\mathbf{f}_{1,0} + \alpha\sqrt{N}\left\|\mathbf{d}_0\right\|,\end{aligned}$$

with equality iff $\mathbf{f}_1-\mathbf{f}_{1,0}$ is proportional to $\mathbf{d}_0$. This in turns holds iff $\mathbf{f}_1$ is given by (17). Now note that the elements of $\mathbf{f}_1$ are positive, so that $\mathbf{f}_1$ also solves the constrained problem.

For $\mathcal{G}_\beta$ the solution is similar - we note that $\mathbf{G}$ given by (18) solves the unconstrained problem and is a non-negative definite matrix.

For $\mathcal{H}_\gamma$, we have that $\left\|\mathbf{B}_0\mathbf{h}\right\|^2$ is maximized, subject to $\mathbf{h}^T\mathbf{h}\leq\gamma\sqrt{N}$, iff $\mathbf{h}$ is an eigenvector of $\mathbf{B}_0^T\mathbf{B}_0$, with $\left\|\mathbf{h}\right\|^2=\gamma\sqrt{N}$, corresponding to the largest eigenvalue $\lambda_{\max}\left(\mathbf{B}_0^T\mathbf{B}_0\right) = \lambda_{\max}\left(\mathbf{B}_0\mathbf{B}_0^T\right)$. But then $\mathbf{B}_0^T\mathbf{B}_0\mathbf{h}=\lambda_{\max}\mathbf{h}$ and so

$$\mathbf{Z}^T\mathbf{h} = \frac{\mathbf{Z}^T\mathbf{B}_0^T\mathbf{B}_0\mathbf{h}}{\lambda_{\max}} = \mathbf{0},$$

since $\mathbf{B}_0\mathbf{Z}=\mathbf{0}$. Thus the orthogonality constraint (2) of $\mathcal{H}_\gamma$ is also satisfied.

Collecting these three maxima gives (16). □

Proof of Lemma 1: (i) First assume A1). Then $\mathbf{Q}_1\mathbf{P}_0\rightarrow\mathbf{H}:=\mathbf{Z}_1\left(\mathbf{Z}_1^T\mathbf{Z}_1\right)^{-1}\mathbf{Z}_1^T$. Disregarding terms which are $o(1)$ we have $\mathbf{Q}_1\left(\mathbf{P}_0-\mathbf{Q}_1^T\mathbf{H}\right)=\mathbf{0}$. Thus the columns of $\mathbf{P}_0-$

$\mathbf{Q}_1^T\mathbf{H}$ are orthogonal to the rows of $\mathbf{Q}_1$, hence are linear combinations of the columns of $\mathbf{Q}_2^T$, i.e. $\mathbf{P}_0 = \mathbf{Q}_1^T\mathbf{H} + \mathbf{Q}_2^T\mathbf{M}$ for some $\mathbf{M}_{(N-n)\times n}$ (necessarily $= \mathbf{Q}_2\mathbf{P}_0$, implying that $\mathbf{MH} = \mathbf{M}$). Thus

$$\mathbf{B}_0 = \mathbf{C}\mathbf{Q}^T \begin{pmatrix} -(\mathbf{I}-\mathbf{H})\,\mathbf{Q}_1 \\ \mathbf{M}\mathbf{Q}_1 - \mathbf{Q}_2 \end{pmatrix}.$$

Note that $(\mathbf{I}-\mathbf{H})\,\mathbf{Q}_1\left(\mathbf{Q}_1^T\mathbf{M}^T - \mathbf{Q}_2^T\right) = \mathbf{0}$. Using this we calculate that

$$\mathbf{B}_0\mathbf{B}_0^T = \mathbf{C}\mathbf{Q}^T\left[(\mathbf{I}-\mathbf{H}) \oplus \left(\mathbf{I}_{N-n} + \mathbf{M}\mathbf{M}^T\right)\right]\mathbf{Q}\mathbf{C}^T.$$

It now follows from either (19) or (20), and the fact that $\mathbf{I}-\mathbf{H}$ has eigenvalues 0 and 1, that $\lambda_{\max}\left(\mathbf{B}_0\mathbf{B}_0^T\right) = 1 + \lambda_{\max}\left(\mathbf{M}^T\mathbf{M}\right)$.

Write $\mathbf{H} = \mathbf{\Lambda}_1\mathbf{\Lambda}_1^T$, where $\mathbf{\Lambda}_1^T\mathbf{\Lambda}_1 = \mathbf{I}_p$. Then

$$\mathbf{P}_0^T\mathbf{P}_0 = \mathbf{H} + \mathbf{M}^T\mathbf{M} = \mathbf{\Lambda}_1\left(\mathbf{I}_p + \mathbf{\Lambda}_1^T\mathbf{M}^T\mathbf{M}\mathbf{\Lambda}_1\right)\mathbf{\Lambda}_1^T$$

has the same non-zero eigenvalues as $\mathbf{I}_p + \mathbf{\Lambda}_1^T\mathbf{M}^T\mathbf{M}\mathbf{\Lambda}_1$, so that the maximum eigenvalue is

$$\begin{aligned} \lambda_{\max}\left(\mathbf{P}_0^T\mathbf{P}_0\right) &= 1 + \lambda_{\max}\left(\mathbf{\Lambda}_1^T\mathbf{M}^T\mathbf{M}\mathbf{\Lambda}_1\right) = 1 + \lambda_{\max}\left(\mathbf{M}^T\mathbf{M}\mathbf{H}\right) \\ &= 1 + \lambda_{\max}\left(\mathbf{M}^T\mathbf{M}\right) = \lambda_{\max}\left(\mathbf{B}_0\mathbf{B}_0^T\right). \end{aligned}$$

(ii) Under A2) we have that $\mathbf{Q}_1\mathbf{P}_0 \to \mathbf{I}_n$ and the remainder of the derivation is very similar to, but simpler than, that under A1). □

ACKNOWLEDGEMENTS

This work has benefited from discussions with Julie Zhou, University of Victoria and Subhash Lele, University of Alberta. The research is supported by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Coakley, C.W., and Hettmansperger, T.P. (1993), "A Bounded Influence, High Breakdown, Efficient Regression Estimator," *Journal of the American Statistical Association*, 88, 872-880.

Cressie, N.A.C., and Hawkins, D.M. (1980), "Robust Estimation of the Variogram I, " *Mathematical Geology*, 12, 115-125.

Cressie, N. (1991), *Statistics for Spatial Data*, New York: Wiley.

Fedorov, Valery V., and Hackl, Peter (1994), "Optimal Experimental Design: Spatial Sampling" *Calcutta Statistical Association Bulletin*, 44, 57-81.

Field, C. A., and Wiens, D.P. (1996), "One-step M-estimators in the Linear Model, With Dependent Errors," *The Canadian Journal of Statistics*, 22, 219-231.

Genton, M.G. (2001), "Robustness Problems in the Analysis of Spatial Data," in: *Spatial Statistics: Methodological Aspects and Application,* ed. M. Moore, New York: Springer, pp. 21 - 37.

Genton, M.G., and Gorsich, D.J. (2002), "Nonparametric Variogram and Covariogram Estimation With Fourier-Bessel Matrices," *Computational Statistics and Data Analysis*, 41, 47-57.

Gomez, M., and Hazen, K. (1970), "Evaluating Sulphur and Ash Distribution in Coal Seams by Statistical Response Surface Regression Analysis," *U.S. Bureau of Mines* Report R1 7377.

Heckman, N.E. (1987), "Robust Design in a Two Treatment Comparison in the Presence of a Covariate," *Journal of Statistical Planning and Inference*, 16, 75-81.

Hill, R.W. (1977), "Robust Regression When There are Outliers in the Carriers," unpublished Ph.D. dissertation, Harvard University, Cambridge, Mass.

Marcus, M.B., and Sacks, J. (1976), "Robust Designs for Regression Problems," in: *Statistical Theory and Related Topics II*, ed. S.S. Gupta and D.S. Moore, New York: Academic Press, pp. 245-268.

Martin, R. J. (1986), "On the Design of Experiments Under Spatial Correlation," (corr: 75, p. 396), *Biometrika*, 73, 247-277.

McArthur, Richard D. (1987), "An Evaluation of Sample Designs for Estimating a Locally Concentrated Pollutant," *Communications in Statistics, Part B – Simulation and Computation*, 16, 735-759.

Militino, A.F., and Ugarte, M.D. (1997), "A GM Estimation of the Location Parameters in a Spatial Linear Model," *Communications in Statistics A*, 26, 1701-1725.

Merrill, H. M., and Schweppe, F. C. (1971), "Bad Data Suppression in Power System Static State Estimation," *IEEE Transactions on Power Applications and Systems*, PAS-90, 2718-2725.

Royle, J.A. (2002), "Exchange Algorithms for Constructing Large Spatial Designs," *Journal of Statistical Planning and Inference*, 100**,** 121-134.

Sacks, J., and Schiller, S. (1988), "Spatial Designs," in *Statistical Decision Theory and Related Topics IV, Volume 2*, 385-395.

Schilling, Mark F. (1992), "Spatial Designs When the Observations are Correlated," *Communications in Statistics, Part B – Simulation and Computation*, 21: 243-267.

Silvapullé, M.J. (1985), "Asymptotic Behavior of Robust Estimators of Regression and Scale Parameters With Fixed Carriers," *The Annals of Statistics*, 13, 1490-1497.

Simpson, D.G., Ruppert, D., and Carroll, R.J. (1992), "On One-Step GM Estimates and Stability of Inferences in Linear Regression," *Journal of the American Statistical Association*, 87, 439-450.

Sinha, S. and Wiens, D.P. (2002), "Robust Sequential Designs for Nonlinear Regression," *The Canadian Journal of Statistics*, 30, 601-618.

Stein, Michael L. (1995),"Locally Lattice Sampling Designs for Isotropic Random Fields," *The Annals of Statistics*, 23, 1991-2012.

Thompson, S. K. (1997), "Effective Sampling Strategies for Spatial Studies," *Metron*, 55**,** 3-21.

Wiens, D.P. (1996), "Asymptotics of Generalized M-Estimation of Regression and Scale With Fixed Carriers, in an Approximately Linear Model," *Statistics and Probability Letters*, 30, 271-285.