The Effects of Common Random Numbers on Stochastic Kriging Metamodels

Xi Chen
Bruce Ankenman
Barry L. Nelson

Department of Industrial Engineering & Management Sciences
Northwestern University

October 24, 2011

Abstract

Ankenman et al. introduced stochastic kriging as a metamodeling tool for representing stochastic simulation response surfaces, and employed a very simple example to suggest that the use of common random numbers (CRN) degrades the capability of stochastic kriging to predict the true response surface. In this paper we undertake an in-depth analysis of the interaction between CRN and stochastic kriging by analyzing a richer collection of models and performing an empirical study. We also consider the effect of CRN on metamodel parameter estimation and response-surface gradient estimation, as well as response-surface prediction. In brief, we confirm that CRN is detrimental to prediction, but show that it leads to better estimation of slope parameters and superior gradient estimation compared to independent simulation.

1 Introduction

Beginning with the seminal papers of Kleijnen (1975) and Schruben and Margolin (1978), simulation researchers have been interested in the impact of incorporating common random numbers (CRN) into experiment designs for fitting linear–regression metamodels of the form

$$Y(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\top \boldsymbol{\beta} + \varepsilon \tag{1}$$

to the output of stochastic simulation experiments. In Model (1), $Y(\mathbf{x})$ is the simulation output, $\mathbf{x} = (x_1, x_2, \ldots, x_p)^\top$ is a vector of controllable design or decision variables, $\mathbf{f}(\mathbf{x})$ is a vector of known functions of $\mathbf{x}$ (e.g., x_1, x_3^2, x_1x_7), $\boldsymbol{\beta}$ is a vector of unknown parameters of appropriate dimension, and ε represents the intrinsic variability in the simulation output assuming no bias in this metamodel.

CRN is a variance reduction technique that attempts to induce a positive correlation between the outputs of simulation experiments at distinct design points (settings of $\mathbf{x}$ in the context of Model (1)) and thereby reduce the variance of the estimator of the expected value of their difference. For $k \geq 2$ design points, a large literature has shown that, properly applied, CRN reduces the variance of "slope" parameters in (1)—and therefore estimates of the response-surface gradient—while often inflating the variance of the intercept term. See, for instance, Donohue et al. (1992, 1995), Hussey et al. (1987ab), Kleijnen (1988, 1992), Nozari et al. (1987), and Tew and Wilson (1992, 1994).

It is fair to say that for Model (1) the role of CRN has been thoroughly examined. The purpose of this paper is to undertake a similar analysis of the interaction of CRN and a new metamodeling technique called stochastic kriging (Ankenman, et al. 2008, 2010). Stochastic kriging is an extension of kriging, which is typically applied to deterministic computer experiments (see, for instance, Santner et al. 2003), to stochastic simulation. Kriging treats the unknown response surface as a realization of a Gaussian random field that exhibits spatial correlation, while stochastic kriging accounts for the additional uncertainty in stochastic simulation due to intrinsic sampling noise. Stochastic kriging is related to kriging with a "nugget effect" that treats the measurement errors as independent and identically distributed mean-zero random variables; stochastic kriging makes modeling additional properties of the random errors possible, namely unequal variances and correlation of the random errors across the design space. The focus of this paper is the effects of introducing correlated random errors via CRN.

Ankenman et al. (2010) used a two-point problem with all parameters known to show that CRN increases the mean squared error (MSE) of the MSE-optimal predictor at a prediction point that has equal spatial correlation with the two design points. They speculated that CRN will not be helpful for prediction in general. In this paper we generalize their two-point problem to allow unequal spatial correlations between the design points and the prediction point, and drop the assumption that the trend-model parameters are known. Therefore we show that the detrimental effect of CRN was not an artifact of the assumptions of Ankenman et al. (2010). We then extend the result given in Appendix EC.2 in Ankenman et al. (2010) for $k \geq 2$ spatially uncorrelated design points and show that CRN inflates the MSE of prediction. We assume that the trend parameters are unknown whereas Ankenman et al. (2010) assumed that all parameters are known. In contrast to prediction, we show that CRN typically improves the estimation of trend-model parameters (i.e., $\boldsymbol{\beta}$) by reducing the variances of the slope parameters; CRN also improves gradient estimation in the sense that the gradient estimators from stochastic kriging are less affected by simulation noise when CRN is employed. A numerical study looks into the joint effect on prediction of using CRN and estimating the intrinsic variance; estimating the intrinsic variance is fundamental to stochastic kriging. All of these results are obtained under the assumption that the parameters of the spatial correlation model are known. Therefore, we close this paper with two empirical studies in which this assumption is relaxed, and we evaluate the effects of CRN on parameter estimation, prediction and gradient estimation in the context of estimating all the parameters of the stochastic kriging model.

2 Stochastic Kriging

In this section we briefly review stochastic kriging as developed in Ankenman et al. (2010) and the particular simplifications we exploit in this paper.

In stochastic kriging we represent the simulation's output on replication j at design point $\mathbf{x}$ as

$$\mathcal{Y}_j(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\top \boldsymbol{\beta} + \mathsf{M}(\mathbf{x}) + \varepsilon_j(\mathbf{x}) = \mathsf{Y}(\mathbf{x}) + \varepsilon_j(\mathbf{x}) \tag{2}$$

where M is a realization of a mean zero *Gaussian random field*; that is, we think of M as being randomly sampled from a space of functions mapping $\Re^p \to \Re$. Therefore, $\mathsf{Y}(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\top \boldsymbol{\beta} + \mathsf{M}(\mathbf{x})$ represents the unknown response surface at point $\mathbf{x}$. In this paper we will focus with only one exception on the special case

$$\mathsf{Y}(\mathbf{x}) = \beta_0 + \sum_{d=1}^{p} \beta_d x_d + \mathsf{M}(\mathbf{x}). \tag{3}$$

Finally, $\varepsilon_1(\mathbf{x}), \varepsilon_2(\mathbf{x}), \ldots$ represents the independent identically distributed sampling noise observed for each replication taken at design point $\mathbf{x}$. We sometimes refer to $\mathsf{M}(\mathbf{x})$ and $\varepsilon_j(\mathbf{x})$ as the extrinsic and intrinsic uncertainties respectively at design point $\mathbf{x}$, as they were defined in Ankenman et al. (2010).

Yin et al. (2010) also propose an extension of kriging to stochastic simulation. Their metamodel is similar to Equation (2), except that $\varepsilon_j(\mathbf{x})$ is also modelled as a Gaussian random field that is independent of M, and they take a fully Bayesian approach by treating all of the model parameters as having prior distributions. While directly accounting for parameter uncertainty, their model does not allow the effect of CRN to be separated from the spatial structure of the intrinsic variance of the simulation output.

For most of the analysis in this paper we assume that the variance $\mathsf{V} = \mathsf{V}(\mathbf{x}) \equiv \mathrm{Var}[\varepsilon(\mathbf{x})]$ at all design points is equal, while allowing the possibility that $\rho(\mathbf{x}, \mathbf{x}') \equiv \mathrm{Corr}[\varepsilon_j(\mathbf{x}), \varepsilon_j(\mathbf{x}')] > 0$ due to CRN. In most discrete-event simulation settings the variance of the intrinsic noise $\mathsf{V}(\mathbf{x})$ depends (perhaps strongly) on the location of design point, $\mathbf{x}$, and one of the key contributions of stochastic kriging is to address experiment design and analysis when this is the case. However, there are a number of reasons that we will not consider heterogeneous intrinsic variance except in the empirical study: In practice, $\mathsf{V}(\mathbf{x})$ can take many forms, making it nearly impossible to obtain useful expressions for the effect of CRN. Further, if the variance of the noise depends on $\mathbf{x}$, then complicated experiment design techniques (e.g., as developed in Ankenman et al. 2010) are needed to properly counteract the effects of the non-constant variance. Once again, this would not lead to tractable results. In some sense, the equal variance assumption used in this paper is intended to represent the conditions after the proper experiment design strategy has mitigated the effects of the non-constant variance. We do include one example in the empirical study that manifests non-constant $\mathsf{V}(\mathbf{x})$ as a check that our conclusions are unaffected.

In our setting an experiment design consists of n simulation replications taken at all k design points $\{\mathbf{x}_i\}_{i=1}^k$. When we assume equal variances, then taking n the same at all design points seems reasonable and again greatly simplifies the analysis; furthermore, equal n is appropriate for CRN so that replication j has a companion for all design points. Let the sample mean of simulation output at $\mathbf{x}_i$ be

$$\begin{aligned} \bar{\mathcal{Y}}(\mathbf{x}_i) &= \frac{1}{n}\sum_{j=1}^{n}\mathcal{Y}_j(\mathbf{x}_i) \\ &= \mathsf{Y}(\mathbf{x}_i) + \frac{1}{n}\sum_{j=1}^{n}\varepsilon_j(\mathbf{x}_i) \qquad (4) \\ &= \beta_0 + \sum_{d=1}^{p}\beta_d x_d + \mathsf{M}(\mathbf{x}_i) + \frac{1}{n}\sum_{j=1}^{n}\varepsilon_j(\mathbf{x}_i) \end{aligned}$$

and let $\bar{\mathcal{Y}} = \left(\bar{\mathcal{Y}}(\mathbf{x}_1), \bar{\mathcal{Y}}(\mathbf{x}_2), \ldots, \bar{\mathcal{Y}}(\mathbf{x}_k)\right)^\top$. Define $\Sigma_{\mathsf{M}}(\mathbf{x}, \mathbf{x}') = \mathrm{Cov}[\mathsf{M}(\mathbf{x}), \mathsf{M}(\mathbf{x}')]$ to be the covariance of points $\mathbf{x}$ and $\mathbf{x}'$ implied by the extrinsic spatial correlation model; and let the $k \times k$ matrix $\mathbf{\Sigma}_{\mathsf{M}}$ be the extrinsic spatial variance-covariance matrix of the k design points $\{\mathbf{x}_i\}_{i=1}^{k}$. Finally, let $\mathbf{x}_0$ be the prediction point, and define $\mathbf{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot)$ to be the $k \times 1$ vector that contains the extrinsic spatial covariances between $\mathbf{x}_0$ and each of the k design points; that is,

$$\mathbf{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot) = \left(\mathrm{Cov}[\mathsf{M}(\mathbf{x}_0), \mathsf{M}(\mathbf{x}_1)], \mathrm{Cov}[\mathsf{M}(\mathbf{x}_0), \mathsf{M}(\mathbf{x}_2)], \ldots, \mathrm{Cov}[\mathsf{M}(\mathbf{x}_0), \mathsf{M}(\mathbf{x}_k)]\right)^\top.$$

Since M is stationary, $\mathbf{\Sigma}_{\mathsf{M}}$ and $\mathbf{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot)$ are of the following form:

$$\mathbf{\Sigma}_{\mathsf{M}} = \tau^2 \begin{pmatrix} 1 & r_{12} & \ldots & r_{1k} \\ r_{21} & 1 & \ldots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & \ldots & 1 \end{pmatrix} \qquad \text{and} \qquad \mathbf{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot) = \tau^2 \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_k \end{pmatrix}$$

where $\tau^2 > 0$ is the extrinsic spatial variance. Gradient estimation only makes sense if the response surface is differentiable. The differentiability of Gaussian process models like in Equation (3) depend on the differentiability of its spatial correlation function as the distance between design points goes to zero. See, for instance, Santer et al. (2003) Section 2.3.4. In particular, the sample paths are infinitely differentiable if the popular Gaussian correlation function is used. Therefore we choose to adopt the Gaussian correlation function $\mathrm{Corr}[\mathsf{M}(\mathbf{x}_i), \mathsf{M}(\mathbf{x}_\ell)] = \exp\{-\sum_{j=1}^{p}\theta_j(x_{ij} - x_{\ell j})^2\}$ in this paper. To simplify notation, the spatial correlation between the design point $\mathbf{x}_i$ and the prediction point $\mathbf{x}_0$ is $r_i = \mathrm{Corr}[\mathsf{M}(\mathbf{x}_0), \mathsf{M}(\mathbf{x}_i)]$, and the spatial correlation between two design points $\mathbf{x}_h$ and $\mathbf{x}_i$ is $r_{hi} = \mathrm{Corr}[\mathsf{M}(\mathbf{x}_h), \mathsf{M}(\mathbf{x}_i)]$. To obtain tractable results, the spatial correlation parameter is assumed to be the same across all dimensions in this paper; i.e., $\theta_j = \theta,\ j = 1, 2, \ldots, p$. This assumption, although not always appropriate in practice, helps facilitate the analysis and demonstrate the theme of this paper without introducing unnecessary technical difficulties.

To make the k-point models tractable, in forthcoming Sections 3.2 and 4.2 we let $\mathbf{\Sigma}_{\mathsf{M}} = \tau^2\,\mathbf{I}_k$ where $\mathbf{I}_k$ denotes the $k \times k$ identity matrix. This form of $\mathbf{\Sigma}_{\mathsf{M}}$ indicates that the design points are spatially uncorrelated with one another, which might be plausible if the design points are widely separated in the region of interest. In addition, to derive results for the k-point trend model in Section 4.2, we further assume that $\mathbf{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot) = \tau^2(r_0, r_0, \ldots, r_0)^\top$; this scenario might be plausible if the design points are widely separated, say at the extremes of the region of interest, while $\mathbf{x}_0$ is central. These assumptions are useful for insight and tractability, but not necessary for stochastic kriging.

What distinguishes stochastic kriging from kriging is that we account for the sampling variability inherent in a stochastic simulation. Let $\mathbf{\Sigma}_\varepsilon$ be the $k \times k$ variance-covariance matrix implied by the sample average intrinsic noise with (h, i) element

$$\mathbf{\Sigma}_\varepsilon(\mathbf{x}_h, \mathbf{x}_i) = \mathrm{Cov}\left[\sum_{j=1}^{n} \varepsilon_j(\mathbf{x}_h)/n, \sum_{j=1}^{n} \varepsilon_j(\mathbf{x}_i)/n\right]$$

across all design points $\mathbf{x}_h$ and $\mathbf{x}_i$. The anticipated effect of CRN is to cause the off-diagonal elements of $\mathbf{\Sigma}_\varepsilon$ to be positive. To make our results tractable in Sections 3 and 4, we let

$$\mathbf{\Sigma}_\varepsilon = \frac{\mathsf{V}}{n}\begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix} \tag{5}$$

where $\rho > 0$, meaning we assume equal variance and correlation. Again, these assumptions are useful for insight and tractability, but not necessary for stochastic kriging. The MSE-optimal predictor (metamodel) provided by stochastic kriging takes the form

$$\widehat{\mathsf{Y}}(\mathbf{x}_0) = \mathbf{f}(\mathbf{x}_0)^\top\widehat{\boldsymbol{\beta}} + \mathbf{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot)^\top[\mathbf{\Sigma}_{\mathsf{M}} + \mathbf{\Sigma}_\varepsilon]^{-1}(\bar{\mathcal{Y}} - \mathbf{F}\widehat{\boldsymbol{\beta}})$$

where the rows of $\mathbf{F}$ are $\mathbf{f}(\mathbf{x}_1)^\top, \mathbf{f}(\mathbf{x}_2)^\top, \ldots, \mathbf{f}(\mathbf{x}_k)^\top$. In the mathematical analysis in Sections 3 and 4, we will suppose that only $\boldsymbol{\beta}$ needs to be estimated, while $\mathbf{\Sigma}_{\mathsf{M}}$, $\mathbf{\Sigma}_\varepsilon$ and $\mathbf{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot)$ are known. In Section 5, we consider what happens when $\mathbf{\Sigma}_\varepsilon$ is estimated, and numerically assess its impact on prediction performance. Finally, our empirical studies in Sections 6 and 7 will estimate every parameter and re-examine the effects of CRN in this context.

3 Intercept Models

In kriging metamodeling for deterministic computer experiments, the most common form is the intercept model (no other trend terms, better known as "ordinary kriging") since (it is argued) the random field term M is flexible enough to account for any variation across the response surface. In this section, we study intercept models and how the use of CRN affects parameter estimation, prediction and gradient estimation. All results are derived in Appendix A.

3.1 A Two-Point Intercept Model

Consider the two-point intercept model $\mathcal{Y}_j(x) = \beta_0 + \mathsf{M}(x) + \varepsilon_j(x)$ with β_0 unknown, design points x_1 and x_2 with equal numbers of replications n, and prediction point x_0, with $x_i \in \Re, i = 0, 1, 2$. Therefore, $\mathsf{Y}(x_0) = \beta_0 + \mathsf{M}(x_0)$ is the response that we want to predict, β_0 is the parameter we want to estimate, and $d\mathsf{Y}(x_0)/dx_0$ is the gradient of interest.

The best linear unbiased predictor (BLUP) of $\mathsf{Y}(x_0)$, the stochastic kriging predictor, is

$$\widehat{\mathsf{Y}}(x_0) = \frac{\bar{\mathcal{Y}}(x_1) + \bar{\mathcal{Y}}(x_2)}{2} + \frac{\tau^2 \left(\frac{\bar{\mathcal{Y}}(x_1) - \bar{\mathcal{Y}}(x_2)}{2}\right)}{\tau^2(1 - r_{12}) + \frac{\mathsf{V}}{n}(1 - \rho)}(r_1 - r_2) \qquad (6)$$

with MSE

$$\text{MSE}^\star = \tau^2\left(1 - (r_1 + r_2)\right) + \frac{1}{2}\left[\tau^2(1 + r_{12}) + \frac{\mathsf{V}}{n}(1 + \rho) - \frac{\tau^4(r_1 - r_2)^2}{\tau^2(1 - r_{12}) + \frac{\mathsf{V}}{n}(1 - \rho)}\right]. \qquad (7)$$

We can show that $d\text{MSE}^\star/d\rho$ is always positive, hence it follows that the use of CRN, which tends to increase ρ, increases the MSE* of the best linear unbiased predictor for this two-point intercept model. Notice that for the spatial variance-covariance matrix of $(\mathsf{Y}(x_0), \bar{\mathcal{Y}}(x_1), \bar{\mathcal{Y}}(x_2))^\top$ to be positive definite, the following condition must be satisfied: $-r_{12}^2 + 2r_1r_2r_{12} + 1 - (r_1^2 + r_2^2) > 0$.

The best linear unbiased estimator (BLUE) of β_0 corresponding to the BLUP of $\mathsf{Y}(x_0)$ is

$$\widehat{\beta}_0 = \frac{\bar{\mathcal{Y}}(x_1) + \bar{\mathcal{Y}}(x_2)}{2} \qquad (8)$$

and it is easy to see that its variance is increasing in ρ since it is a sum of positively correlated outputs. Thus, the MSE of prediction and the variance of $\widehat{\beta}_0$ are both inflated by CRN.

Let $\widehat{\nabla}_{\text{sk}}$ denote the gradient of the predictor $\widehat{\mathsf{Y}}(x_0)$ at x_0 in the stochastic kriging setting. Under the assumptions given in Section 2, it follows that

$$\begin{aligned}\widehat{\nabla}_{\text{sk}} &= \frac{d\widehat{\mathsf{Y}}(x_0)}{dx_0} \\ &= -2\theta\left[r_1(x_0 - x_1) + r_2(x_2 - x_0)\right] \frac{\tau^2 \left(\frac{\bar{\mathcal{Y}}(x_1) - \bar{\mathcal{Y}}(x_2)}{2}\right)}{\left[\tau^2(1 - r_{12}) + \frac{\mathsf{V}}{n}(1 - \rho)\right]}. \qquad (9)\end{aligned}$$

To assess the impact of CRN, we choose as a benchmark the gradient estimator that would be obtained if there were no simulation intrinsic variance; that is, if the response surface could be observed noise free. We are interested in the impact of CRN on the "distance" between the noisy and noise-free gradient estimators to measure whether CRN helps mitigate the effect of intrinsic variance on gradient estimation.

Let $\widehat{\nabla}_{\text{sk}}(n)$ be the gradient estimator when n simulation replications are used at each design point, and let $\widehat{\nabla}_{\text{sk}}(\infty)$ be the gradient estimator as $n \to \infty$, which can be obtained by simply setting the intrinsic variance $\mathsf{V} = 0$ in Equation (9). It follows that

$$\mathrm{E}\Big[\widehat{\nabla}_{\mathrm{sk}}(n) - \widehat{\nabla}_{\mathrm{sk}}(\infty)\Big]^2 = \frac{2\theta^2(r_1(x_0 - x_1) + r_2(x_2 - x_0))^2}{\big((1 - r_{12})/[\frac{\mathsf{V}}{n}(1 - \rho)] + 1/\tau^2\big)(1 - r_{12})}. \tag{10}$$

From Equation (10), we see that CRN decreases the mean squared difference between these two estimators. In the extreme case as $\rho \to 1$, even if n is not large, the gradient estimator from stochastic kriging converges to the "ideal" case because the effect of stochastic noise on gradient estimation is eliminated by employing CRN.

3.2 A k-Point Intercept Model

In the previous section we were able to show that CRN is detrimental to response surface prediction and parameter estimation, but is beneficial to gradient estimation in a two-design-point setting. In this section we are able to draw the same conclusions in a particular k-point ($k \geq 2$) intercept model, $\mathcal{Y}_j(\mathbf{x}) = \beta_0 + \mathsf{M}(\mathbf{x}) + \varepsilon_j(\mathbf{x})$, with β_0 unknown. Under the assumptions given in Section 2, the following results can be obtained:

The BLUP of $\mathsf{Y}(\mathbf{x}_0)$ is

$$\widehat{\mathsf{Y}}(\mathbf{x}_0) = \frac{1}{k}\sum_{i=1}^{k}\bar{\mathcal{Y}}(\mathbf{x}_i) + \frac{\tau^2}{\left(\frac{\mathsf{V}}{n}(1-\rho) + \tau^2\right)}\left(\sum_{i=1}^{k} r_i\bar{\mathcal{Y}}(\mathbf{x}_i) - \frac{1}{k}\left(\sum_{i=1}^{k}\bar{\mathcal{Y}}(\mathbf{x}_i)\right)\left(\sum_{i=1}^{k} r_i\right)\right) \tag{11}$$

with MSE

$$\begin{aligned}\mathrm{MSE}^{\star} &= \tau^2 + \frac{\tau^4}{\frac{\mathsf{V}}{n}(1-\rho) + \tau^2}\left(\frac{1}{k}\left(\sum_{i=1}^{k} r_i\right)^2 - \sum_{i=1}^{k} {r_i}^2\right) \\ &\quad + \frac{\frac{\mathsf{V}}{n}\left((k-1)\rho + 1\right) + \tau^2}{k} - 2\tau^2\left(\frac{1}{k}\sum_{i=1}^{k} r_i\right). \end{aligned} \tag{12}$$

Notice that for the spatial variance-covariance matrix of $(\mathsf{Y}(\mathbf{x}_0), \bar{\mathcal{Y}}(\mathbf{x}_1), \ldots, \bar{\mathcal{Y}}(\mathbf{x}_k))^\top$ to be positive definite, it must be that $\sum_{i=1}^{k} r_i^2 < 1$. We show in Appendix A that under this condition $d\mathrm{MSE}^{\star}/d\rho$ is positive for any $\rho \in [0, 1)$, hence CRN increases $\mathrm{MSE}^{\star}$.

The BLUE of β_0 corresponding to the BLUP of $\mathsf{Y}(\mathbf{x}_0)$ is

$$\widehat{\beta}_0 = \frac{1}{k}\sum_{i=1}^{k}\bar{\mathcal{Y}}(\mathbf{x}_i) \tag{13}$$

and its variance is easily shown to be an increasing function of ρ.

Similar to the analysis of gradient estimation in Section 3.1, let $\widehat{\nabla}_{\mathrm{sk}} = \left(\widehat{\nabla}_{\mathrm{sk}_1}, \widehat{\nabla}_{\mathrm{sk}_2}, \ldots, \widehat{\nabla}_{\mathrm{sk}_p}\right)^\top$ denote the gradient of $\widehat{\mathsf{Y}}(\mathbf{x}_0)$ at $\mathbf{x}_0$ in the stochastic kriging setting; notice that now $\widehat{\nabla}_{\mathrm{sk}}$

is a random vector in $\Re^p$. We can show that for $j = 1, 2, \ldots, p$, the jth component of the gradient is

$$\begin{aligned} \widehat{\nabla}_{\mathrm{sk}_j} &= \frac{\partial \widehat{\mathsf{Y}}(\mathbf{x}_0)}{\partial x_{0j}} \\ &= \frac{-2\theta\tau^2}{\tau^2 + \frac{\mathsf{V}}{n}(1-\rho)} \cdot \sum_{i=1}^{k} \left(\left(\bar{\mathcal{Y}}(\mathbf{x}_i) - \frac{1}{k}\sum_{h=1}^{k} \bar{\mathcal{Y}}(\mathbf{x}_h) \right) (x_{0j} - x_{ij}) r_i \right) \end{aligned} \qquad (14)$$

where the ith design point $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$ is a vector in $\Re^p$, $i = 1, \ldots, k$. Recall that $r_i = \exp\{-\theta \sum_{j=1}^{p} (x_{0j} - x_{ij})^2\}$ is the spatial correlation between $\mathbf{x}_i$ and $\mathbf{x}_0$, and that we assume that the design points are spatially uncorrelated, meaning that they are separated enough that $r_{ij} \approx 0$, for $i \neq j$.

Now for $p > 2$, we continue to use $\widehat{\nabla}_{\mathrm{sk}}(\infty)$ as the benchmark to evaluate gradient estimation in the stochastic kriging setting. We use the inner product to measure the "distance" between the two random vectors $\widehat{\nabla}_{\mathrm{sk}}(n)$ and $\widehat{\nabla}_{\mathrm{sk}}(\infty)$ at prediction point $\mathbf{x}_0 \in \Re^p$ and call it the mean squared difference between these two gradient estimators. We can show that

$$\begin{aligned} \langle \widehat{\nabla}_{\mathrm{sk}}(n) - \widehat{\nabla}_{\mathrm{sk}}(\infty), \widehat{\nabla}_{\mathrm{sk}}(n) - \widehat{\nabla}_{\mathrm{sk}}(\infty) \rangle &= \sum_{j=1}^{p} \mathrm{E}\left[\left(\widehat{\nabla}_{\mathrm{sk}_j}(n) - \widehat{\nabla}_{\mathrm{sk}_j}(\infty) \right)^2 \right] \\ = \frac{4\theta^2}{\left(\frac{1}{\frac{\mathsf{V}}{n}(1-\rho)} + \frac{1}{\tau^2} \right)} \sum_{j=1}^{p} \left(\sum_{i=1}^{k} (x_{0j} - x_{ij})^2 r_i^2 - \frac{1}{k} \left(\sum_{i=1}^{k} (x_{0j} - x_{ij}) r_i \right)^2 \right). & \end{aligned} \qquad (15)$$

As in Section 3.1, we arrive at the conclusion that for this k-point intercept model, CRN decreases the mean squared difference between these two gradient estimators.

4 Trend Models

Although many practitioners use intercept models for kriging, it remains to be seen what models will be most effective when noise is introduced. Also, in linear regression models, CRN is known to be most helpful for estimating slope parameters and so it seems likely that CRN will perform best under a trend model that, like a regression model, includes slope parameters. For these reasons and for completeness, we next study the effects of CRN on stochastic kriging with a trend model (a counterpart of " universal kriging" in kriging context).

4.1 A Two-Point Trend Model

Consider the two-point trend model $\mathcal{Y}_j(x) = \beta_0 + \beta_1 x + \mathsf{M}(x) + \varepsilon_j(x)$ with β_0 and β_1 unknown, so that $\mathsf{Y}(x_0) = \beta_0 + \beta_1 x_0 + \mathsf{M}(x_0)$ is the unknown response that we want to predict at point x_0. Without loss of generality, suppose that $x_1 < x_2$. Then we can show the following results:

The BLUP of $\mathsf{Y}(x_0)$ is

$$\widehat{\mathsf{Y}}(x_0) = \frac{\bar{\mathcal{Y}}(x_2)(x_0 - x_1) + \bar{\mathcal{Y}}(x_1)(x_2 - x_0)}{(x_2 - x_1)} \tag{16}$$

with MSE

$$\mathrm{MSE}^\star = 2\tau^2 + \frac{\mathsf{V}}{n} - \frac{2ab}{(a+b)^2}\left[\tau^2(1 - r_{12}) + \frac{\mathsf{V}}{n}(1-\rho)\right] - \frac{2\tau^2}{(a+b)}(ar_1 + br_2) \tag{17}$$

where $a = x_2 - x_0$, $b = x_0 - x_1$, $a + b = x_2 - x_1$. Equation (17) implies that for this two-point trend model, when $x_0 \in (x_1, x_2)$, CRN increases MSE*; however, if we do extrapolation, i.e., $x_0 \notin (x_1, x_2)$, then CRN will decrease MSE*. Notice that the existing literature on kriging claims that kriging does not perform well in extrapolation, so kriging should be restricted to interpolation. Finally, if $x_0 = x_1$ or x_2, we get $\widehat{\mathsf{Y}}(x_0) = \bar{\mathcal{Y}}(x_1)$ or $\bar{\mathcal{Y}}(x_2)$, respectively; in this case MSE* is reduced to V/n, the same with and without using CRN.

The BLUE of $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$ corresponding to the BLUP of $\mathsf{Y}(x_0)$ is

$$\widehat{\boldsymbol{\beta}} = \frac{1}{(x_2 - x_1)}\begin{pmatrix} x_2\bar{\mathcal{Y}}(x_1) - x_1\bar{\mathcal{Y}}(x_2) \\ \bar{\mathcal{Y}}(x_2) - \bar{\mathcal{Y}}(x_1) \end{pmatrix}. \tag{18}$$

It follows that

$$\mathrm{Var}(\widehat{\beta}_0) = \left(\tau^2 + \frac{\mathsf{V}}{n}\right) + \frac{2x_1x_2}{(x_2 - x_1)^2}\left[\tau^2(1 - r_{12}) + \frac{\mathsf{V}}{n}(1-\rho)\right] \tag{19}$$

$$\mathrm{Var}(\widehat{\beta}_1) = \frac{2\left[\tau^2(1 - r_{12}) + \frac{\mathsf{V}}{n}(1-\rho)\right]}{(x_2 - x_1)^2} \tag{20}$$

and

$$\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = \frac{-(x_1 + x_2)\left[\tau^2(1 - r_{12}) + \frac{\mathsf{V}}{n}(1-\rho)\right]}{(x_2 - x_1)^2}.$$

From Equation (20), we see that CRN reduces the variance of $\widehat{\beta}_1$. Also notice that Equation (19) implies that if $x_1x_2 < 0$, so that 0 is interior to the design space, then CRN inflates the variance of $\widehat{\beta}_0$, while if $x_1x_2 > 0$, so that $\widehat{\beta}_0$ is an extrapolated prediction of the response at $x = 0$, then CRN decreases the variance of $\widehat{\beta}_0$.

Finally, following the analysis in Section 3.1, we can show that the mean squared difference between the gradient estimators obtained when the number of replications n is finite and when $n \rightarrow \infty$ is

$$\mathrm{E}\left[\widehat{\nabla}_{\mathrm{sk}}(n) - \widehat{\nabla}_{\mathrm{sk}}(\infty)\right]^2 = \frac{2\mathsf{V}(1-\rho)}{n(x_1 - x_2)^2}. \tag{21}$$

Equation (21) shows that CRN decreases the mean squared difference between these two estimators. Observe that the extrinsic spatial variance τ^2 has no influence on this mean squared difference at all.

4.2 A k-Point Trend Model

For the two-point trend model we were able to draw conclusions similar to those we found for the intercept model and an additional conclusion related to the estimation of the slope parameter. Specifically, we found that CRN is detrimental to response surface prediction at any point *inside the region of experimentation* since it increases the MSE of prediction, but CRN is beneficial to estimation of the slope parameter by decreasing the variance of its estimator and beneficial to gradient estimation since it decreases the effect of noise. As with the intercept model we can extend the conclusions of the two-point trend model to a k-point ($k \geq 2$) trend model if additional restrictions are made.

Consider the k-point trend model $\mathcal{Y}_j(\mathbf{x}) = \beta_0 + \sum_{d=1}^{p} \beta_d x_d + \mathsf{M}(\mathbf{x}) + \varepsilon_j(\mathbf{x})$, where $p \geq 2$. Suppose that we have a $k \times (p+1)$ *orthogonal* design matrix $\mathbf{D}_k$ of rank $(p+1)$

$$\mathbf{D}_k = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{k1} & \dots & x_{kp} \end{pmatrix}$$

which means that the column vectors of $\mathbf{D}_k$ are pairwise orthogonal. Such an assumption on $\mathbf{D}_k$ is not common for kriging, because kriging usually employs space-filling designs such as a Latin Hypercube Sample; nevertheless in addition to the assumptions given in Section 2 orthogonality makes the analysis tractable enough to give the following results:

The BLUE of $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^\top$ corresponding to the BLUP of $\mathsf{Y}(\mathbf{x}_0)$ is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{D}_k^\top \boldsymbol{\Sigma}^{-1} \mathbf{D}_k)^{-1} \mathbf{D}_k^\top \boldsymbol{\Sigma}^{-1} \bar{\mathcal{Y}} \tag{22}$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\mathsf{M}} + \boldsymbol{\Sigma}_{\varepsilon}$. More explicitly,

$$\widehat{\beta}_0 = \frac{1}{k} \sum_{i=1}^{k} \bar{\mathcal{Y}}(\mathbf{x}_i) \tag{23}$$

and

$$\widehat{\beta}_j = \frac{\sum_{i=1}^{k} x_{ij} \bar{\mathcal{Y}}(\mathbf{x}_i)}{\sum_{i=1}^{k} x_{ij}^2}, \quad j = 1, 2, \ldots, p. \tag{24}$$

The resulting BLUP of $\mathsf{Y}(\mathbf{x}_0)$ is

$$\widehat{\mathsf{Y}}(\mathbf{x}_0) = \mathbf{f}(\mathbf{x}_0)^\top \widehat{\boldsymbol{\beta}} \tag{25}$$

where $\mathbf{f}(\mathbf{x}_0) = (1, x_{01}, x_{02}, \ldots, x_{0p})^\top$. The corresponding optimal MSE is

$$\begin{aligned} \mathrm{MSE}^\star \quad = \quad & \tau^2 \left(1 + \frac{1}{k} + \sum_{j=1}^{p} \frac{x_{0j}^2}{\sum_{i=1}^{k} x_{ij}^2} - 2 r_0 \right) + \frac{1}{k} \frac{\mathsf{V}}{n} \left(1 + k \sum_{j=1}^{p} \frac{x_{0j}^2}{\sum_{i=1}^{k} x_{ij}^2} \right) \\ & + \frac{1}{k} \frac{\mathsf{V}}{n} \rho \left((k-1) - k \sum_{j=1}^{p} \frac{x_{0j}^2}{\sum_{i=1}^{k} x_{ij}^2} \right). \end{aligned} \tag{26}$$

Notice that if

$$\frac{k-1}{k} > \sum_{j=1}^{p} \frac{x_{0j}^2}{\sum_{i=1}^{k} x_{ij}^2} \tag{27}$$

then CRN increases MSE*.

To help interpret this result, consider a $k = 2^p$ factorial design where the design points are $x_{ij} \in \{-1, +1\}$. Then Equation (27) reduces to $\sum_{j=1}^{p} x_{0j}^2 < k-1$. Therefore, CRN will inflate the MSE* of $\widehat{\mathsf{Y}}(\mathbf{x}_0)$ at prediction points inside a sphere of radius $\sqrt{2^p - 1}$ centered at the origin (which is also the center of the experiment design). Notice that for $p > 1$ we have $\sqrt{2^p - 1} > \sqrt{p}$, the radius of the sphere that just contains the design points and is the usual prediction region of interest. Also notice that when $p = 1$ we recover the condition for the two-point trend model, for which we have more general results available in Section 4.1 without the orthogonality assumption.

We next focus on the effect of CRN on $\mathrm{Cov}(\widehat{\boldsymbol{\beta}})$. Because of the orthogonality assumption, the expression for $\mathrm{Cov}(\widehat{\boldsymbol{\beta}})$ becomes much simpler. It can be shown that

$$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = (\mathbf{D}_k^\top \boldsymbol{\Sigma}^{-1} \mathbf{D}_k)^{-1} = \begin{pmatrix} \frac{\frac{\mathsf{V}}{n}[1+(k-1)\rho]+\tau^2}{k} & 0 & \cdots & 0 \\ 0 & \frac{\frac{\mathsf{V}}{n}(1-\rho)+\tau^2}{\sum_{i=1}^{k} x_{i1}^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\frac{\mathsf{V}}{n}(1-\rho)+\tau^2}{\sum_{i=1}^{k} x_{ip}^2} \end{pmatrix}. \tag{28}$$

Hence we arrive at a similar conclusion to the one obtained in Section 4.1: CRN reduces the variances of $\widehat{\beta}_1, \widehat{\beta}_2, \cdots, \widehat{\beta}_p$. Here the first diagonal term manifests that CRN increases $\mathrm{Var}(\widehat{\beta}_0)$, which is consistent with Section 4.1 since $\mathbf{0}$ is interior to the design space.

Now let

$$\widehat{\nabla}_{\mathrm{sk}} = (\widehat{\nabla}_{\mathrm{sk}_1}, \widehat{\nabla}_{\mathrm{sk}_2}, \ldots, \widehat{\nabla}_{\mathrm{sk}_p})^\top$$

denote the gradient of $\widehat{\mathsf{Y}}(\mathbf{x}_0)$ at $\mathbf{x}_0$ in the stochastic kriging setting. We can show that for $j = 1, 2, \ldots, p$, the jth component of the gradient is

$$\begin{aligned} \widehat{\nabla}_{\mathrm{sk}_j} &= \frac{\partial \widehat{\mathsf{Y}}(\mathbf{x}_0)}{\partial x_{0j}} \\ &= \frac{d\boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot)}{dx_{0j}} \boldsymbol{\Sigma}^{-1}(\bar{\mathcal{Y}} - \mathbf{D}\widehat{\boldsymbol{\beta}}) + \widehat{\beta}_j \\ &= \widehat{\beta}_j. \end{aligned}$$

Following the analysis in Section 3.2, we define the following inner product to measure the "distance" between the two random vectors $\widehat{\nabla}_{\text{sk}}(n)$ and $\widehat{\nabla}_{\text{sk}}(\infty)$ at prediction point $\mathbf{x}_0$:

$$\begin{aligned}\langle \widehat{\nabla}_{\text{sk}}(n) - \widehat{\nabla}_{\text{sk}}(\infty), \widehat{\nabla}_{\text{sk}}(n) - \widehat{\nabla}_{\text{sk}}(\infty)\rangle &= \sum_{j=1}^{p} \mathrm{E}\left[\left(\widehat{\nabla}_{\text{sk}_j}(n) - \widehat{\nabla}_{\text{sk}_j}(\infty)\right)^2\right] \\ &= \frac{\mathsf{V}}{n}(1-\rho)\sum_{j=1}^{p}\left(\sum_{i=1}^{k} x_{ij}^2\right)^{-1}. \qquad (29)\end{aligned}$$

Equation (29) shows that CRN decreases the mean squared difference between these two gradient estimators. Similar to the result in Section 4.1, we see that only the intrinsic noise affects this mean squared difference, whereas the extrinsic spatial variance has no influence on it at all.

5 Estimating the Intrinsic Variance-Covariance Matrix

A key component of stochastic kriging is estimating $\boldsymbol{\Sigma}_\varepsilon$. In this section, we first present a theorem stating that estimating the intrinsic variance-covariance matrix will not lead to biased prediction. Then by modifying the k-point intercept model in Section 3.2, we proceed to study the impact of estimating the common intrinsic variance with correlated random noise among design points induced by CRN. In the following analysis, we assume that $\boldsymbol{\Sigma}_\varepsilon$ is unknown but everything else is known including $\boldsymbol{\beta}$.

We begin with formally stating the following assumption:

Assumption 1 *The random field* M *is a stationary Gaussian random field. For design point* $\mathbf{x}_i$*,* $i = 1, 2, \ldots, k$*, the random noise from different replications* $\varepsilon_1(\mathbf{x}_i), \varepsilon_2(\mathbf{x}_i), \ldots$ *are i.i.d.* $\mathcal{N}(0, \mathsf{V}(\mathbf{x}))$*. For the* j*th replication,* $j = 1, 2, \ldots, n$*, the* $k \times 1$ *vector of random noise across all design points* $[\varepsilon_j(\mathbf{x}_1), \varepsilon_j(\mathbf{x}_2), \ldots, \varepsilon_j(\mathbf{x}_k)]^\top$ *has a multivariate normal distribution with mean* $\mathbf{0}$ *and variance-covariance matrix* $\widetilde{\boldsymbol{\Sigma}}_\varepsilon$ *(with usage of CRN). The random noise is independent of* M*.*

Notice that the only new condition added given those already stated in Section 2 is the multivariate normal distribution of the random noise across all design points in the same simulation replication. Under Assumption 1, $(\mathsf{Y}(\mathbf{x}_0), \bar{\mathcal{Y}}(\mathbf{x}_1), \ldots, \bar{\mathcal{Y}}(\mathbf{x}_k))$ is multivariate normally distributed, which follows from a proof similar to Ankenman et al. (2010). The stochastic kriging predictor (41) given at the beginning of Appendix A.2 is the conditional expectation of $\mathsf{Y}(\mathbf{x}_0)$ given $\bar{\mathcal{Y}}$. The $k \times 1$ vector of sample average random noise at all k design points $[\bar{\varepsilon}(\mathbf{x}_1), \bar{\varepsilon}(\mathbf{x}_2), \ldots, \bar{\varepsilon}(\mathbf{x}_k)]^\top$ has a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}_\varepsilon$, where $\bar{\varepsilon}(\mathbf{x}_i) = n^{-1}\sum_{j=1}^{n} \varepsilon_j(\mathbf{x}_i),\ i = 1, 2, \ldots, k$ and $\varepsilon_j(\mathbf{x}_i)$ is the random noise at design point $\mathbf{x}_i$ in the jth replication. It follows that $\boldsymbol{\Sigma}_\varepsilon = n^{-1}\widetilde{\boldsymbol{\Sigma}}_\varepsilon$.

Now let $\mathcal{S}$ denote the sample variance-covariance matrix of the intrinsic noise across the k design points:

$$\mathcal{S} = \begin{pmatrix} \mathcal{S}_{11} & \mathcal{S}_{12} & \dots & \mathcal{S}_{1k} \\ \mathcal{S}_{21} & \mathcal{S}_{22} & \dots & \mathcal{S}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{S}_{k1} & \mathcal{S}_{k2} & \dots & \mathcal{S}_{kk} \end{pmatrix} \tag{30}$$

where

$$\begin{aligned} \mathcal{S}_{i\ell} &= \frac{1}{n-1}\sum_{j=1}^{n} (\mathcal{Y}_j(\mathbf{x}_i) - \bar{\mathcal{Y}}(\mathbf{x}_i))(\mathcal{Y}_j(\mathbf{x}_\ell) - \bar{\mathcal{Y}}(\mathbf{x}_\ell)) \qquad (31)\\ &= \frac{1}{n-1}\sum_{j=1}^{n} (\varepsilon_j(\mathbf{x}_i) - \bar{\varepsilon}(\mathbf{x}_i))(\varepsilon_j(\mathbf{x}_\ell) - \bar{\varepsilon}(\mathbf{x}_\ell)) \ . \end{aligned}$$

In words, $\mathcal{S}_{i\ell}$ is the sample covariance of the random noise at design points $\mathbf{x}_i$ and $\mathbf{x}_\ell$, $i, \ell = 1, 2, \dots, k$. We use $n^{-1}\mathcal{S}$ to estimate $\mathbf{\Sigma}_\varepsilon$. The next result shows that estimating $\mathbf{\Sigma}_\varepsilon$ in this way introduces no prediction bias. The proof can be found in Appendix A.8.

Theorem 1 *Let $\widehat{\mathbf{\Sigma}}_\varepsilon = n^{-1}\mathcal{S}$, where $\mathcal{S}$ is specified as in Equation* (31). *Define*

$$\widehat{\widehat{\mathsf{Y}}}(\mathbf{x}_0) = \mathbf{f}(\mathbf{x}_0)^\top \boldsymbol{\beta} + \mathbf{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot)^\top \left[\mathbf{\Sigma}_{\mathsf{M}} + \widehat{\mathbf{\Sigma}}_\varepsilon\right]^{-1} (\bar{\mathcal{Y}} - \mathbf{F}\boldsymbol{\beta}) \tag{32}$$

where $\mathbf{f}(\mathbf{x}_i)$ denotes the $(q+1) \times 1$ vector of functions $\mathbf{f}(\mathbf{x}_i), i = 0, 1, \dots, k$ and $\mathbf{F}$ is the $k \times (q+1)$ design matrix of full rank

$$\mathbf{F} = \begin{pmatrix} \mathbf{f}(\mathbf{x}_1)^\top \\ \mathbf{f}(\mathbf{x}_2)^\top \\ \vdots \\ \mathbf{f}(\mathbf{x}_k)^\top \end{pmatrix} .$$

If Assumption 1 holds, then $\mathrm{E}\left[\widehat{\widehat{\mathsf{Y}}}(\mathbf{x}_0) - \mathsf{Y}(\mathbf{x}_0)\right] = 0.$

Recall the k-point intercept model $\mathcal{Y}_j(\mathbf{x}) = \beta_0 + \mathsf{M}(\mathbf{x}) + \varepsilon_j(\mathbf{x})$, where the design points $\{\mathbf{x}_i\}_{i=1}^k$ are in $\Re^p$ and equal numbers of replications n are obtained from each of them. In Ankenman et al. (2010) the effect of estimating intrinsic variance was investigated assuming the intrinsic noise at each design point to be independent and identically distributed with a common intrinsic variance. Following Ankenman et al. (2010), we next focus on how much variance inflation occurs when $\mathbf{\Sigma}_\varepsilon$ is estimated under the same assumptions as in Ankenman et al. (2010) but with the addition of CRN. Suppose

$$\mathbf{\Sigma}_{\mathsf{M}} = \tau^2 \begin{pmatrix} 1 & r & \cdots & r \\ r & 1 & \cdots & r \\ \vdots & \vdots & \ddots & \vdots \\ r & r & \cdots & 1 \end{pmatrix}$$

and $\mathbf{\Sigma}_{\mathsf{M}}(\mathbf{x}_0,\cdot) = \tau^2(r_0, r_0, \ldots, r_0)^\top$ with $r_0, r \geq 0$. This represents a situation in which the extrinsic spatial correlations among the design points are all equal and the design points are equally correlated with the prediction point.

Notice that for the spatial variance-covariance matrix of $(\mathsf{Y}(\mathbf{x}_0), \bar{\mathcal{Y}}(\mathbf{x}_1), \ldots, \bar{\mathcal{Y}}(\mathbf{x}_k))^\top$ to be positive definite, the condition $r_0^2 < 1/k + r(k-1)/k$ must be satisfied. To make the analysis tractable but still interesting, we assume that $\widetilde{\mathbf{\Sigma}}_\varepsilon$ has the following form, with ρ known and V unknown:

$$\widetilde{\mathbf{\Sigma}}_\varepsilon = \mathsf{V}\begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix}.$$

Hence it follows that $\mathbf{\Sigma}_\varepsilon = n^{-1}\widetilde{\mathbf{\Sigma}}_\varepsilon$. As in Ankenman et al. (2010), we suppose that there is an estimator $\widehat{\mathsf{V}}$ of V such that $\widehat{\mathsf{V}} \sim \mathsf{V}\chi^2_{n-1}/(n-1)$, viz. $(n-1)\widehat{\mathsf{V}}/\mathsf{V}$ has a chi-squared distribution with degrees of freedom $n-1$. In Appendix A.9 we show that the MSE of $\widehat{\mathsf{Y}}(\mathbf{x}_0)$, the stochastic kriging predictor with V known, is

$$\mathrm{MSE}^\star = \tau^2\left(1 - \frac{kr_0^2}{1 + C_{\rho\gamma} + (k-1)r}\right) \tag{33}$$

where $C_{\rho\gamma} = \frac{\gamma}{n}(1 + (k-1)\rho)$ and $\gamma = \mathsf{V}/\tau^2$ denotes the ratio of the intrinsic variance to the extrinsic variance, which is (roughly speaking) a measure of the sampling noise relative to the response surface variation. On the other hand, the MSE of $\widehat{\widehat{\mathsf{Y}}}(\mathbf{x}_0)$ obtained by substituting $\widehat{\mathsf{V}}$ for V is

$$\mathrm{MSE} = \tau^2\mathrm{E}\left[1 + \frac{kr_0^2\left[1 + C_{\rho\gamma} + (k-1)r\right]}{\left(1 + \frac{\widehat{\mathsf{V}}}{\mathsf{V}}C_{\rho\gamma} + (k-1)r\right)^2} - \frac{2kr_0^2}{\left(1 + \frac{\widehat{\mathsf{V}}}{\mathsf{V}}C_{\rho\gamma} + (k-1)r\right)}\right]. \tag{34}$$

We assess the MSE inflation and the effect of CRN on it by evaluating the ratio of (34) to (33) numerically. The MSE inflation ratio is largest when n is small and r_0 and r are large, so in the numerical analysis we show the inflation ratio as a function of $\gamma = \mathsf{V}/\tau^2$ and ρ for $n = 10$, $r = 0, 0.1, 0.2$ and r_0 at 95% of the maximum value it can take. We use $k = 50$ design points throughout the study for convenient comparison with the results given in Ankenman et al. (2010).

We summarize our findings as follows and refer readers to Appendix A.10 for a detailed discussion. There is a penalty associated with estimating intrinsic variance; that is, doing so always inflates prediction MSE relative to using the (unknown) true value of $\mathbf{\Sigma}_\varepsilon$. However, *for a fixed value of the spatial correlation* of a given response surface, CRN can either magnify or diminish this penalty depending on the ratio of the intrinsic variance to the extrinsic variance, or in other words, depending on which source of variation dominates for that particular response surface. The MSE inflation that results from estimating $\mathbf{\Sigma}_\varepsilon$ is even more substantial in the presence of CRN when spatial variation dominates intrinsic variation ($\tau^2 \gg \mathsf{V}$). On the other hand, the MSE inflation from estimating $\mathbf{\Sigma}_\varepsilon$ is dimishished by using CRN when intrinsic variation dominates spatial variation ($\tau^2 \ll \mathsf{V}$).

These effects of CRN on MSE inflation hold for response surfaces with varying degrees of smoothness. Interestingly, we found that *how smooth a given response surface is* in turn leads an overall "controlling" role. Specifically, strong spatial correlation of the response surface tends to counteract the effect of CRN on MSE inflation, whatever it is. A response surface with "strong spatial correlation" tends to be smoother than one with weaker spatial dependence, since the value of the response at any point tends to be similar to—that is, strongly correlated with—other points in close proximity. When CRN magnifies the MSE inflation, then strong spatial correlation reduces the magnification. On the other hand, when CRN diminishes the MSE inflation, it is less effective at doing so when the surface exhibits strong spatial correlation.

Lastly we suggest that discretion needs to be exercised when one interprets the results above, because even if the MSE inflation ratio is close to 1, the MSE* itself can be large; therefore ratio = 1 does not mean that the particular setting provides a good prediction. Similarly, a large MSE inflation ratio does not necessarily imply that a particular experimental setting provides poor prediction. Finally from the discussion in Appendix A.10 we conclude that even with this small value of n (recall that $n = 10$), the MSE inflation ratio is slight over an extreme range of $\gamma = \mathsf{V}/\tau^2$. As n increases, the inflation vanishes. This suggests that the penalty for estimating V will typically be small.

6 An Experiment with Gaussian Random Fields

From the two-point and k-point intercept and trend models we gained some insight into the impact of CRN on parameter estimation, prediction and gradient estimation for stochastic kriging. However, to obtain these results we had to assume all model parameters except $\boldsymbol{\beta}$ (Sections 3–4) and $\mathbf{\Sigma}_\varepsilon$ (Section 5) were known. In this section, we confirm these insights empirically when all parameters must be estimated. The factors we investigate are the strength of the correlation ρ induced by CRN; the number of design points k; the strength of the extrinsic spatial correlation coefficient θ; and the ratio of the intrinsic variance to the extrinsic variance $\gamma = \mathsf{V}/\tau^2$.

We consider a one-dimensional problem where the true response surface is $\mathsf{Y}(x) = 10 + 3x + \mathsf{M}(x)$ with $x \in [0, 1]$. The Gaussian random field M, denoted by $\mathrm{GRF}(\tau^2, \theta)$, has extrinsic spatial covariance between points x and x' given by $\Sigma_{\mathsf{M}}(x, x') = \tau^2 \exp\{-\theta(x - x')^2\}$. A test-function instance is created by sampling $\mathsf{M} \sim \mathrm{GRF}(\tau^2, \theta)$, and we sample multiple instances as part of the experiment. We fix $\tau^2 = 1$ but θ is varied to obtain smooth and rough response-surface instances.

The simulation response observed at point x on replication j is $\mathcal{Y}_j(x) = 10 + 3x + \mathsf{M}(x) + \varepsilon_j(x)$, where the random noise $\varepsilon_j(x),\ j = 1, 2, \ldots, n$ is i.i.d. $\mathcal{N}(0, \mathsf{V})$; since we assume equal variance it is reasonable to take the same number of replications, n, at each design point. The effect of CRN is represented by specifying a common correlation $\rho = \mathrm{Corr}[(\varepsilon_j(x), \varepsilon_j(x')]$

for $x \neq x', j = 1, 2, \ldots, n$. We vary $\gamma = \mathsf{V}/\tau^2 = \mathsf{V}$ to introduce random noise of different relative intensities.

An equally spaced grid design of k design points $x \in [0, 1]$ is used, with $k \in \Omega_k = \{4, 7, 13, 25\}$. We make $n = 100$ replications at each design point, and control V so that $\gamma/n = \mathsf{V}/n \in \Omega_\gamma = \{0.01, 0.25, 1\}$, corresponding to low, medium and high intrinsic variance. We took $\theta \in \Omega_\theta = \{4.6052, 13.8155\}$ (or equivalently, $\exp(-\theta) \in \{0.01, 10^{-6}\}$) ; small θ tends to give a smoother response surface. We vary ρ in $\Omega_\rho = \{0, 0.4, 0.8\}$ to assess the effect of increasing correlation induced by CRN; For each $\theta \in \Omega_\theta$ we sample 10 true response surfaces, and for each response surface we run 5 macroreplications for each $\{k, \rho, \gamma\} \in \Omega_k \times \Omega_\rho \times \Omega_\gamma$ combination; for a fixed $\{\theta, k, \rho, \gamma\}$, the macroreplications differ only in their randomly sampled $\varepsilon_j(x)$.

Thus, altogether there are $2 \times 10 \times 4 \times 3 \times 3 \times 5 = 3600$ experiments. For each one we fit a stochastic kriging metamodel using maximum likelihood estimation as described in Ankenman et al. (2010), do prediction and gradient estimation at 193 equally spaced points in $[0, 1]$, and record the values of the estimated parameters $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\tau^2}$ and $\widehat{\theta}$. The stochastic kriging code used in these experiments can be found at `www.stochastickriging.net`.

We evaluate the impact on prediction by MSSE, the mean of the sum of squared errors of the predicted values at the 193 check points, namely, $\mathrm{MSSE}(\widehat{\mathsf{Y}}) = 193^{-1} \sum_{i=1}^{193} (\mathsf{Y}(\mathbf{x}_i) - \widehat{\mathsf{Y}}(\mathbf{x}_i))^2$; we evaluate parameter estimation by recording the absolute difference between the true and estimated parameter on each trial; and we evaluate gradient estimation by computing the sample correlation between the true and estimated gradient across the 193 check points.

A brief preview of our findings is as follows.

- CRN does not aid prediction and instead increases the MSSE.
- CRN does reduce the variability of the slope estimator $\widehat{\beta}_1$.
- CRN does improve gradient estimation in the sense that it introduces a strong positive correlation between the estimated gradient and the true gradient.

These findings are consistent with our results in the previous sections.

Boxplots in Figures 1–4 provide more details. For brevity, we show only graphs corresponding to the number of design points $k = 7$. In each figure, the left panel shows the sample statistics of interest obtained from the smoother response surface with $\theta = 4.6052$; while the right panel shows the statistics obtained from the rougher response surface with $\theta = 13.8155$. Within each panel from the left to the right, 3 groups of boxplots are ordered by increasing γ/n; within each group, three individual boxplots are ordered by increasing ρ. Notice that each individual boxplot is a summary of 50 data points from 5 macro-replications on each of 10 surfaces.

To evaluate prediction, we calculate the MSSE by averaging the squared difference between the true response $\mathsf{Y}(x_0)$ and the predicted value $\widehat{\mathsf{Y}}(x_0)$ at x_0 across 193 check points. A summary of the MSSE for $k = 7$ is shown in Figure 1. It is easy to see that increasing ρ increases MSSE, and leads to wider interquantile range. This is especially true when θ is large, or equivalently, when the extrinsic spatial correlation is small. As we expected,

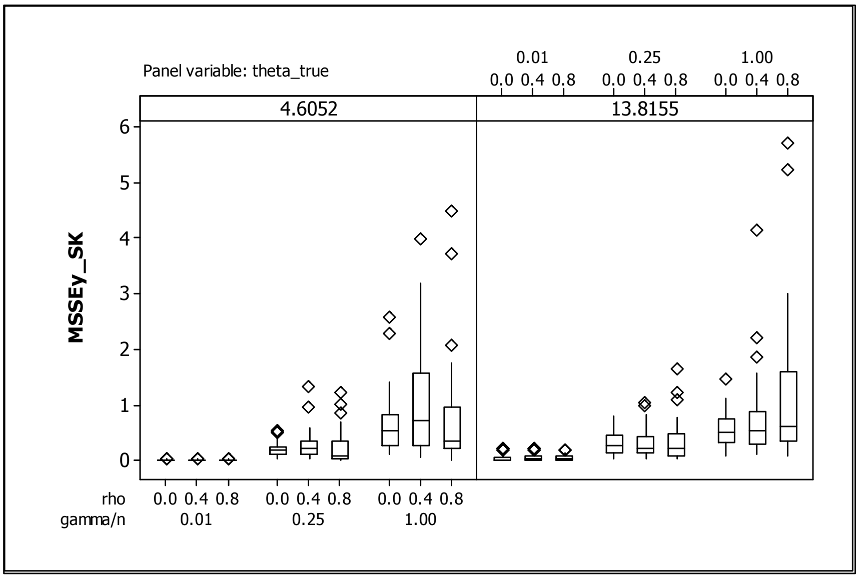

Figure 1: MSSE for $k = 7$.

for fixed ρ, increasing γ/n will increase MSSE. On the other hand, we mention (without showing graphs) that for fixed ρ, increasing the number of design points k leads to narrower interquantile range when γ/n is not large. Finally by observing the sets of three boxplots that are grouped close together to show the effect of increasing ρ, we conclude that CRN does not help prediction.

For parameter estimation, we use the absolute deviation (AD), that is, $|\beta_j - \widehat{\beta}_j|$. A summary of the statistical dispersion of $|\beta_j - \widehat{\beta}_j|$, $j = 0$ and 1 for $k = 7$ is shown in Figures 2 and 3. For $|\beta_1 - \widehat{\beta}_1|$, we see in Figure 3 that increasing ρ decreases $|\beta_1 - \widehat{\beta}_1|$; this effect is more evident when θ is small. The effect of ρ on $|\beta_0 - \widehat{\beta}_0|$ is not as obvious as on $|\beta_1 - \widehat{\beta}_1|$. As we expected, for fixed ρ, increasing γ/n leads to increased ADs and wider interquantile range for both parameters. Finally, we mention (without showing graphs) that increasing the number of design points k moves the interquantile range closer to 0 and helps to estimate slope parameter even better. We conclude that CRN improves estimate of the slope parameter, but its effect on the intercept parameter is not as clear.

To evaluate gradient estimation, we use the correlation between the true gradient and the gradient estimated in the stochastic kriging setting instead of using the mean squared difference between them, since in most applications it is far more important to find the correct direction of change rather than the magnitude of this change. Therefore, $\mathrm{Corr}(\widehat{\nabla}_{\mathrm{true}}(n), \widehat{\nabla}_{\mathrm{sk}}(n))$ gives a better view of the effect of ρ on gradient estimation under the influence of θ and k. We use the finite-difference gradient estimate from the noiseless response data as the true gradient $\widehat{\nabla}_{\mathrm{true}}(n)$. A summary of the correlations between $\widehat{\nabla}_{\mathrm{sk}}(n)$ and $\widehat{\nabla}_{\mathrm{true}}(n)$ for $k = 7$ is shown in Figure 4. It is obvious that increasing ρ consistently increases $\mathrm{Corr}(\widehat{\nabla}_{\mathrm{true}}(n), \widehat{\nabla}_{\mathrm{sk}}(n))$ for all γ/n values, and makes the interquantile range narrower as well

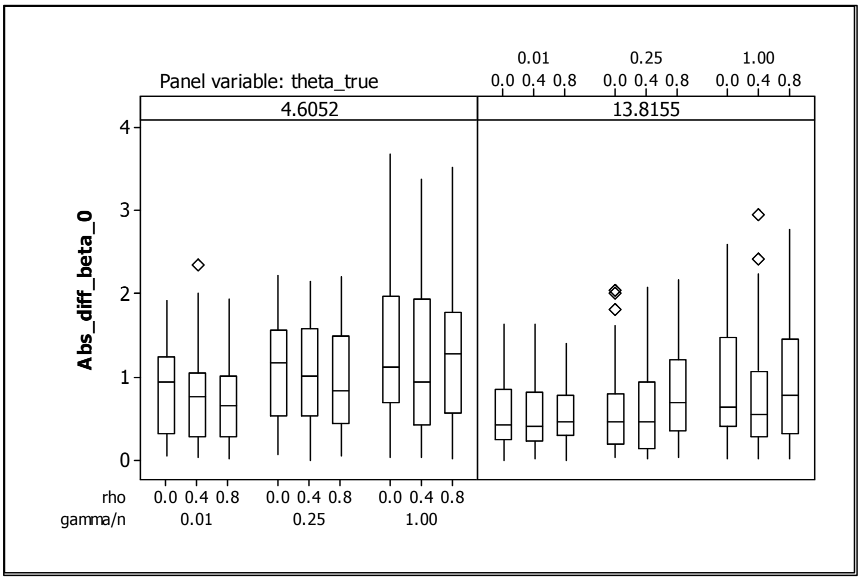

Figure 2: $|\beta_0 - \widehat{\beta_0}|$ for $k = 7$.

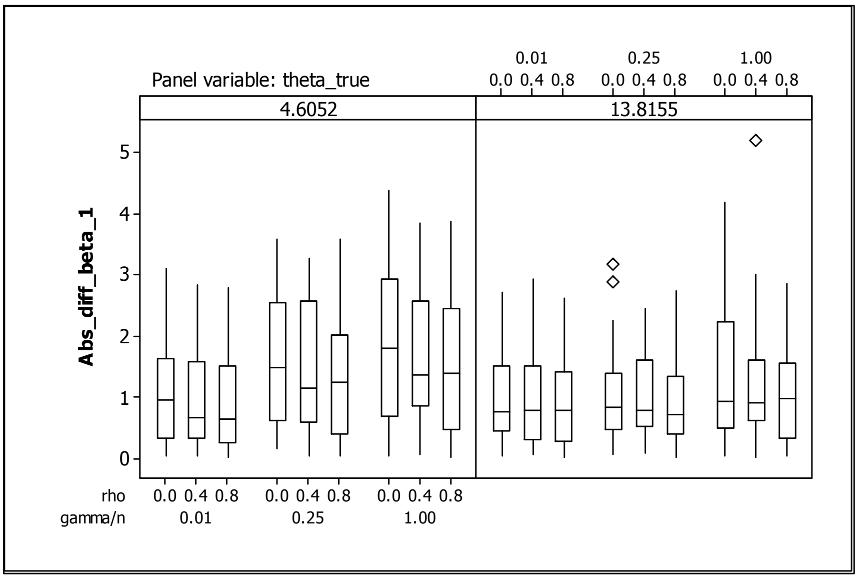

Figure 3: $|\beta_1 - \widehat{\beta_1}|$ for $k = 7$.

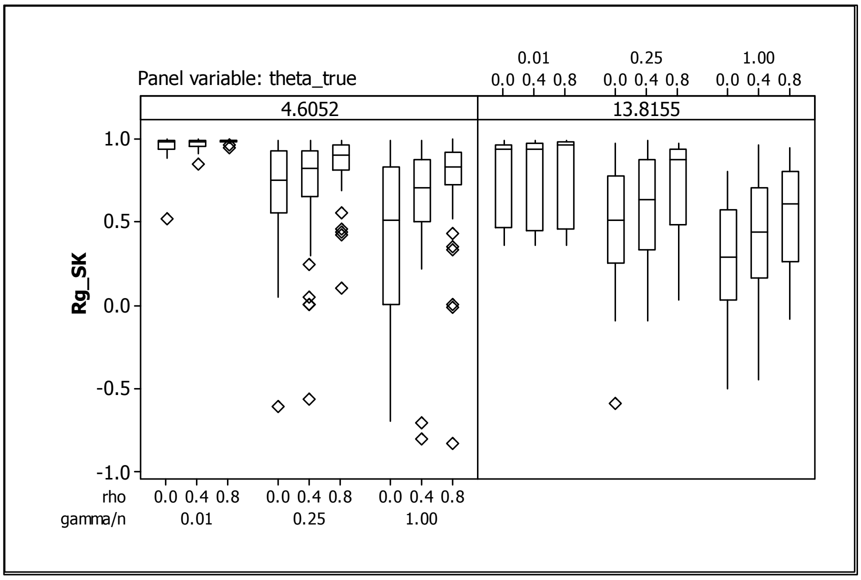

Figure 4: $\mathrm{Corr}(\widehat{\nabla}_{\text{true}}(n),\widehat{\nabla}_{\text{sk}}(n))$ for $k = 7$.

as moving them toward 1; in fact, this effect is more manifest when γ/n is large. Typically, for fixed ρ, increasing γ/n decreases $\mathrm{Corr}(\widehat{\nabla}_{\text{true}}(n),\widehat{\nabla}_{\text{sk}}(n))$ and leads to wider interquantile range. Furthermore, we mention (without showing graphs) that increasing the number of design points k increases $\mathrm{Corr}(\widehat{\nabla}_{\text{true}}(n),\widehat{\nabla}_{\text{sk}}(n))$ and makes the interquantile range narrower. We conclude that CRN improves gradient estimation by introducing a strong positive correlation between the estimated gradient and the true gradient.

For each parameter set $\{\rho, \gamma, \theta, k\}$, we also estimated τ^2 and θ for the 10 response surfaces, each with 5 macro-replications. The estimates $\widehat{\theta}$ and $\widehat{\tau^2}$ obtained are not as good as $\widehat{\beta}_0$ and $\widehat{\beta}_1$ when compared to their known true values. For brevity, we choose not to present them here.

7 M/M/∞ Queue Simulation

In this section, we move a step closer to realistic system simulation problems. Let $\mathsf{Y}(\mathbf{x})$ be the expected steady-state number of customers in an M/M/∞ queue with arrival rate x_1 and mean service time x_2; it is known that $\mathsf{Y}(\mathbf{x}) = x_1x_2$ and the distribution of the steady-state queue-length is Poisson with mean $\mathsf{Y}(\mathbf{x})$. Notice that the variance of the response is x_1x_2 which changes across the design space.

Therefore, given values for x_1 and x_2 we can simulate a steady-state observation by generating a Poisson variate with mean x_1x_2. Given a set of design points $\{x_{i1}, x_{i2}\}_{i=1}^{k'}$, we induce correlation across design points by using the inverse CDF method (Law and Kelton 2000), where k' denotes the number of design points used. Specifically, for replication j

$$\mathcal{Y}_j(\mathbf{x}_i) = \mathrm{F}_{\mathbf{x}_i}^{-1}(\mathcal{U}_j), \quad i = 1, 2, \ldots, k' \tag{35}$$

where $\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_n$ are *i.i.d.* U(0,1); n is the number of simulation replications and $\mathrm{F}_{\mathbf{x}_i}^{-1}(\cdot)$ represents the inverse CDF of a Poisson distribution with mean $x_{i1}x_{i2}$. Notice that our experiment differs from what would occur in practice because we only take a single observation of the queue length on each replication, rather than the average queue length over some period of time. This allows us to compute the correlation induced by CRN, values of which typically are greater than 0.9 in this experiment.

In stochastic kriging when the response surface $\mathsf{Y}(\mathbf{x})$ is unknown, we assume that it takes the form $\mathsf{Y}(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\top\boldsymbol{\beta} + \mathsf{M}(\mathbf{x})$. Three different specifications of $\mathbf{f}(\mathbf{x})^\top\boldsymbol{\beta}$ are considered to further evaluate the effects of CRN; they are

Model 1: An intercept-only model, $\mathbf{f}(\mathbf{x})^\top\boldsymbol{\beta} = \beta_0$

Model 2: A misspecified trend model, $\mathbf{f}(\mathbf{x})^\top\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Model 3: A correctly specified trend model, $\mathbf{f}(\mathbf{x})^\top\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$.

By "correctly specified" we mean that Model 3 can recover the true response surface x_1x_2 while the other two cannot.

Our experiment design is as follows. We consider the design space $1 \le x_d \le 5, d = 1, 2$. For design points we use a Latin Hypercube Sample of $k \in \{5, 20, 50\}$ points, and augment the design with the four corner points $(1,1), (1,5), (5,1)$ and $(5,5)$ to avoid extrapolation. Thus, there are $k' = k + 4$ design points in total. At each design point $n = 400$ simulation replications are made either using CRN as in Equation (35), or sampled independently. We then fit stochastic kriging metamodels with trend terms specified as Models 1, 2 and 3 and make 100 macroreplication of the entire experiment.

For each model specification, we evaluate the impact on predication by $\widehat{\mathrm{MISE}}(\widehat{\mathsf{Y}})$, the approximated mean integrated squared error of $\widehat{\mathsf{Y}}$. We evaluate gradient estimation by computing $\widehat{\mathrm{MISE}}(\widehat{\nabla}_{\mathrm{sk}_d})$, $d = 1$ and 2. In both cases, we approximate MISE by using a 2500 check point grid in $[1,5]^2$. Formally,

$$\widehat{\mathrm{MISE}}(\widehat{\mathsf{Y}}) = \frac{1}{100}\sum_{\ell=1}^{100}\frac{1}{2500}\sum_{i=1}^{2500}(\mathsf{Y}(\mathbf{x}_i') - \widehat{\mathsf{Y}}_\ell(\mathbf{x}_i'))^2$$

and

$$\widehat{\mathrm{MISE}}(\widehat{\nabla}_{\mathrm{sk}_d}) = \frac{1}{100}\sum_{\ell=1}^{100}\frac{1}{2500}\sum_{i=1}^{2500}(\nabla_d(\mathbf{x}_i') - \widehat{\nabla}_{\mathrm{sk}_d}(\mathbf{x}_i', \ell))^2.$$

where the integrated squared error (ISE) is approximated by averaging the sum of squared errors over the 2500 check points and the mean integrated squared error (MISE) is approximated by averaging the approximated ISE over 100 macroreplications. Notice that for better presentation of the results, values shown in Tables 1 and 2 are calculated without the scaling factor 1/2500. Finally we give summary statistics for the parameter estimates of the correctly specified trend model (Model 3).

Table 1: The mean integrated squared error of predictions obtained for the three response models with and without using CRN.

	Model 1		Model 2		Model 3	
k'	Indep.	CRN	Indep.	CRN	Indep.	CRN
9	82.5 (9.8)	68 (7)	501 (112)	267 (99)	12 (1)	13 (2)
24	20 (1)	55 (9)	25 (2)	65 (10)	6.2 (0.4)	13 (2)
54	9.5 (0.6)	72 (11)	12 (1)	73 (10)	4.0 (0.3)	7.0 (1.2)

Table 2: The mean integrated squared error of gradient estimates obtained for the three response models with and without using CRN.

	Model 1				Model 2				Model 3			
k'	Indep.		CRN		Indep.		CRN		Indep.		CRN	
	$\frac{\partial\widehat{\mathsf{Y}}(\mathbf{x}_0)}{\partial x_{01}}$	$\frac{\partial\widehat{\mathsf{Y}}(\mathbf{x}_0)}{\partial x_{02}}$	$\frac{\partial\widehat{\mathsf{Y}}(\mathbf{x}_0)}{\partial x_{01}}$	$\frac{\partial\widehat{\mathsf{Y}}(\mathbf{x}_0)}{\partial x_{02}}$	$\frac{\partial\widehat{\mathsf{Y}}(\mathbf{x}_0)}{\partial x_{01}}$	$\frac{\partial\widehat{\mathsf{Y}}(\mathbf{x}_0)}{\partial x_{02}}$	$\frac{\partial\widehat{\mathsf{Y}}(\mathbf{x}_0)}{\partial x_{01}}$	$\frac{\partial\widehat{\mathsf{Y}}(\mathbf{x}_0)}{\partial x_{02}}$	$\frac{\partial\widehat{\mathsf{Y}}(\mathbf{x}_0)}{\partial x_{01}}$	$\frac{\partial\widehat{\mathsf{Y}}(\mathbf{x}_0)}{\partial x_{02}}$	$\frac{\partial\widehat{\mathsf{Y}}(\mathbf{x}_0)}{\partial x_{01}}$	$\frac{\partial\widehat{\mathsf{Y}}(\mathbf{x}_0)}{\partial x_{02}}$
9	84 (10)	90 (15)	8.4 (1.4)	8.8 (1.5)	748 (204)	782 (209)	388 (177)	367 (169)	3.7 (1.1)	3.4 (0.4)	0.48 (0.07)	0.50 (0.08)
24	30 (3)	31 (4)	2.8 (0.4)	2.8 (0.3)	39 (4)	64 (15)	3.2 (0.3)	3.4 (0.4)	2.2 (0.4)	3.4 (0.6)	0.37 (0.05)	0.39 (0.05)
54	15 (2)	19 (3)	3.0 (0.5)	2.9 (0.4)	23 (4)	26 (3)	2.9 (0.4)	2.9 (0.4)	6.7 (2.1)	13 (3)	0.26 (0.04)	0.29 (0.05)

The effects of CRN on prediction, gradient estimation and parameter estimation can be found in Tables 1–3. In each table, the values are averaged over 100 macroreplications; the values in parentheses are the corresponding standard errors. In brief, we found that the results derived in the previous sections still hold; that is, CRN improves gradient estimation and estimation of slope parameters, but does not aid prediction.

In Table 1, we observe for all three model specifications that the $\widehat{\mathrm{MISE}}(\widehat{\mathsf{Y}})$ is smaller with independent sampling than with CRN, with the exception of Models 1 and 2 when the number of design points is very small ($k' = 9$); increasing the number of design points make this effect even more apparent. Notice that the misspecified trend model (Model 2) gives even worse prediction results than the intercept model (Model 1), while the correctly specified trend model is much better.

In Table 2, it is observed that CRN improves gradient estimation for all three response models. The values for the sample means and standard errors obtained on the correctly specified trend model are much smaller than the corresponding values from the other two response models. We observe once again that the misspecified trend model (Model 2) gives much worse gradient estimates than the intercept model (Model 1) does.

Lastly, we are interested in knowing how CRN affects estimates of the slope parameters for the correctly specified trend model. The results given in Table 3 manifest that CRN

Table 3: Results for the slope parameters for the correctly specified trend model with and without using CRN.

	$\widehat{\beta}_0(\beta_0 = 0)$		$\widehat{\beta}_1(\beta_1 = 0)$		$\widehat{\beta}_2(\beta_2 = 0)$		$\widehat{\beta}_3(\beta_3 = 1)$	
k'	Indep.	CRN	Indep.	CRN	Indep.	CRN	Indep.	CRN
9	-0.009 (0.0095)	-0.003 (0.0024)	-0.001 (0.0040)	-0.001 (0.0009)	0.001 (0.0043)	-0.001 (0.0009)	1.000 (0.0016)	1.000 (0.0003)
24	-0.016 (0.0090)	-0.004 (0.0028)	0.004 (0.0031)	-0.001 (0.0007)	0.006 (0.0036)	-0.001 (0.0007)	0.998 (0.0012)	1.000 (0.0002)
54	0.008 (0.0082)	-0.006 (0.0021)	-0.003 (0.0030)	0.000 (0.0005)	-0.004 (0.0029)	0.000 (0.0005)	1.001 (0.0010)	1.000 (0.0002)

reduces variances of the slope parameter estimates to a great extent; increasing the number of design points does not improve the results much in this case. Notice that if the correctly specified trend model is assumed, one is able to successfully recover the true response model with moderately large number of replications.

8 Conclusions

CRN is one of the most widely used variance reduction techniques; in fact, with most simulation software one would have to carefully program the simulation to avoid using it. Therefore, it is important to understand its effect on a new metamodeling technique such as stochastic kriging. Previous research with other metamodels, such as linear regression, has shown that CRN often leads to more precise parameter estimation, especially with slope parameters that are essentially gradients. However, since CRN can inflate the variability of parameters such as the intercept, it can reduce the precision of the actual prediction.

The parameters, the form, and even the underlying assumptions of stochastic kriging are substantially different from traditional metamodels. Nevertheless, in this paper we have provided compelling evidence that CRN has effects on the stochastic kriging metamodel that are similar or at least analogous to the effects seen in more traditional metamodel settings. Specifically, we have used a variety of tractable models to show that CRN leads to 1) less precise prediction of the response surface in terms of MSE, 2) better estimation of the slope terms in any trend model, and 3) better gradient estimation.

In addition, we are able to show that, under a reasonably mild assumption (Assumption 1 in Section 5), estimating the intrinsic variance-covariance matrix $\boldsymbol{\Sigma}_\varepsilon$ introduces no prediction bias to the plug-in BLUP. A thorough numerical analysis of the MSE inflation that is induced by estimating $\boldsymbol{\Sigma}_\varepsilon$ revealed that stronger spatial correlation counteracts the effect of CRN on MSE inflation.

Finally, through an experiment with Gaussian random fields and an M/M/∞ queue example we assessed the impact of CRN on prediction, parameter estimation and gradient estimation when the parameters of the trend model $\boldsymbol{\beta}$, of the random field τ^2 and $\boldsymbol{\theta}$, and the intrinsic variance-covariance matrix $\boldsymbol{\Sigma}_\varepsilon$ are all estimated as would be required in actual application. The conclusions given by the empirical evaluation were consistent with our analytical results.

The implications of our results are that when the actual prediction values matter, CRN is not recommended. Such scenarios might occur in financial risk analysis or tactical decision making where the primary purpose of the metamodel is to produce predictions of the response in places where no actual simulations have been made and the predicitons are needed quickly (long before actual simulation runs would finish). CRN is recommended for use in gradient estimation for simulation optimization or if the metamodel is a physics-based model where better parameter estimates are of great value to, say, establish sensitivities. Sensitivity analysis is particularly useful for verification and validation of simulation models. Since CRN substantially improves the performance of stochastic kriging gradient estimators, a fruitful area for future research is applying stochastic kriging and CRN to simulation optimization.

Acknowledgment

We thank the referees and editors for comments and suggestions for improvement of our paper. This paper is based upon work supported by the National Science Foundation under Grant No. CMMI-0900354. Portions of this paper were previously published in Chen, Ankenman and Nelson (2010).

References

Ankenman, B. E., B. L. Nelson and J. Staum. 2008. Stochastic kriging for simulation metamodeling. *Proceedings of the 2008 Winter Simulation Conference*, 362–370.

Ankenman, B. E., B. L. Nelson and J. Staum. 2010. Stochastic kriging for simulation metamodeling. *Operations Research* **58**, 371–382.

Chen, X., B. Ankenman and B. L. Nelson. 2010. Common Random Numbers and Stochastic Kriging. *Proceedings of the 2010 Winter Simulation Conference*, 947–956.

Donohue, J. M., R. C. Houck and R. H. Myers. 1992. Simulation designs for quadratic response surface models in the presence of model misspecification. *Management Science* **38**, 1765–1791.

Donohue, J. M., E. C. Houck and R. H. Myers. 1995. Simulation designs for the estimation of quadratic response surface gradients in the presence of model misspecification. *Management Science* **41**, 244–262.

Hussey, J. R., R. H. Myers, and E. C. Houck. 1987a. Pseudorandom number assignment in quadratic response surface designs. *IIE Transactions* **19**, 395–403.

Hussey, J. R., R. H. Myers, E. C. Houck. 1987b. Correlated simulation experiments in first-order response surface design. *Operations Research* **35**, 744–758.

Kleijnen, J. P. C. 1975. Antithetic variates, common random numbers and optimal computer time allocation in simulation. *Management Science* **21**, 1176-1185

Kleijnen, J. P. C. 1988. Analyzing simulation experiments with common random numbers. *Management Science* **34**, 65–74.

Kleijnen, J. P. C. 1992. Regression metamodels for simulation with common random numbers: Comparison of validation tests and confidence intervals. *Management Science* **38**, 1164–1185.

Nozari, A., S. F. Arnold and C. D. Pegden. 1987. Statistical analysis for use with the Schruben and Margolin correlation induction strategy. *Operations Research* **35**, 127–139.

Santner, T. J., B. J. Williams and W. I. Notz. 2003. *The Design and Analysis of Computer Experiments*. Springer, NY.

Schruben, L. W., and B. H. Margolin. 1978. Pseudorandom number assignment in statistically designed simulation and distribution sampling experiments. *Journal of the American Statistical Association* **73**, 504–525.

Tew, J. D., and J. R. Wilson. 1992. Validation of simulation analysis methods for the Schruben-Margolin correlation-induction strategy. *Operations Research* **40**, 87–103.

Tew, J. D., and J. R. Wilson. 1994. Estimating simulation metamodels using combined correlation-based variance reduction techniques. *IIE Transactions* **26**, 2–16.

Yin, J., S. H. Ng., K. M. Ng. 2010. A Bayesian metamodeling approach for stochastic simulations. *Proceedings of the 2010 Winter Simulation Conference*, 1055–1066.

A Appendix: Proofs (intended as an online supplement)

A.1 Stochastic Kriging: The General Results

We begin by proving results for the general model considered in Equation (2) under the assumption that $\mathbf{\Sigma}_{\mathsf{M}}$, $\mathbf{\Sigma}_{\varepsilon}$ and $\mathbf{\Sigma}_{\mathsf{M}}(\mathbf{x},\cdot)$ are known, but $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_q)^\top$ is unknown. Most of these results parallel similar derivations in Stein (1999) for kriging, to which the reader can refer for missing details.

The central idea of stochastic kriging is similar to that of kriging, that is, to give prediction of $\mathsf{Y}(\mathbf{x}_0)$ based on the sample averages of the responses $(\bar{\mathcal{Y}}(\mathbf{x}_1), \bar{\mathcal{Y}}(\mathbf{x}_2) \ldots, \bar{\mathcal{Y}}(\mathbf{x}_k))^\top$ at the k design points. We assume that $\mathbf{x}_i \in \Re^p, i = 1, 2, ..., k$. To facilitate explanation, we introduce further notation. Let $\mathbf{f}(\mathbf{x})$ denote a $(q+1) \times 1$ vector of known functions of $\mathbf{x}$. The sample mean of simulation output at $\mathbf{x}_i$ can be written as

$$\bar{\mathcal{Y}}(\mathbf{x}_i) = \mathbf{f}(\mathbf{x}_i)^\top \boldsymbol{\beta} + \mathsf{M}(\mathbf{x}_i) + \bar{\varepsilon}(\mathbf{x}_i).$$

We further define the $k \times (q+1)$ design matrix $\mathbf{F}$ of full rank as follows:

$$\mathbf{F} = \begin{pmatrix} \mathbf{f}(\mathbf{x}_1)^\top \\ \mathbf{f}(\mathbf{x}_2)^\top \\ \vdots \\ \mathbf{f}(\mathbf{x}_k)^\top \end{pmatrix} \tag{36}$$

Now we consider to get the BLUP of $\mathsf{Y}(x_0)$ that takes the form $\lambda_0(\mathbf{x}_0) + \boldsymbol{\lambda}(\mathbf{x}_0)^\top \bar{\mathcal{Y}}$, where $\lambda_0(\mathbf{x}_0)$ and $\boldsymbol{\lambda}(\mathbf{x}_0)$ are weights that depend on $\mathbf{x}_0$ and will be chosen to give the minimum MSE for predicting $\mathsf{Y}(\mathbf{x}_0) = \mathbf{f}(\mathbf{x}_0)^\top \boldsymbol{\beta} + \mathsf{M}(\mathbf{x}_0)$. To simplify the notation we drop the dependence of λ_0 and $\boldsymbol{\lambda}$ on $\mathbf{x}_0$. Therefore we arrive at the following minimization problem

$$\begin{aligned} &\min_{\lambda_0, \boldsymbol{\lambda}} \mathrm{E}\left[(\mathsf{Y}(\mathbf{x}_0) - \lambda_0 - \boldsymbol{\lambda}^\top \bar{\mathcal{Y}})^2\right] \\ &\quad \text{subject to:} \\ &\quad \mathrm{E}[\lambda_0 + \boldsymbol{\lambda}^\top \bar{\mathcal{Y}}] = \mathrm{E}[\mathsf{Y}(\mathbf{x}_0)] \quad \text{for } \boldsymbol{\beta} \in \mathbb{R}^{q+1}. \end{aligned}$$

To solve this problem, first notice that the unbiased constraint is equivalent to $\lambda_0 + \boldsymbol{\lambda}^\top \mathbf{F} \boldsymbol{\beta} = \mathbf{f}(\mathbf{x}_0)^\top \boldsymbol{\beta}$ for all $\boldsymbol{\beta} \in \mathbb{R}^{q+1}$, or $\lambda_0^\star = 0$ and $\mathbf{F}^\top \lambda^\star = \mathbf{f}(\mathbf{x}_0)$. Following a similar line of argument as Stein (1999, Section 1.5), we can show that for $\boldsymbol{\lambda}^\top \bar{\mathcal{Y}}$ to be a BLUP of $\mathsf{Y}(\mathbf{x}_0)$ there must exist a $(q+1) \times 1$ vector $\boldsymbol{\mu}$ such that $(\mathbf{\Sigma}_{\mathsf{M}} + \mathbf{\Sigma}_{\varepsilon})\boldsymbol{\lambda} - \mathbf{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot) = -\mathbf{F}\boldsymbol{\mu}$. To summarize, $\boldsymbol{\lambda}^\top \bar{\mathcal{Y}}$ is a BLUP of $\mathsf{Y}(\mathbf{x}_0)$ if the following condition holds

$$\begin{pmatrix} \mathbf{\Sigma}_{\mathsf{M}} + \mathbf{\Sigma}_{\varepsilon} & \mathbf{F} \\ \mathbf{F}^\top & \mathbf{0}_{q+1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\mu} \end{pmatrix} = \begin{pmatrix} \mathbf{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot) \\ \mathbf{f}(\mathbf{x}_0) \end{pmatrix} \tag{37}$$

where $\mathbf{0}_{q+1}$ denotes the $(q+1)\times(q+1)$ matrix of zeros. From Equation (37) it follows that $\boldsymbol{\lambda}^\star = \left(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{F}(\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F})^{-1}\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0,\cdot) + \boldsymbol{\Sigma}^{-1}\mathbf{F}(\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F})^{-1}\mathbf{f}(\mathbf{x}_0)$, where we let $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\mathsf{M}} + \boldsymbol{\Sigma}_{\varepsilon}$. From $\lambda_0^\star$ and $\boldsymbol{\lambda}^\star$, the BLUP of $\mathsf{Y}(\mathbf{x}_0)$ immediately follows

$$\widehat{\mathsf{Y}}(\mathbf{x}_0) = \boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0,\cdot)^\top\left[\boldsymbol{\Sigma}_{\mathsf{M}} + \boldsymbol{\Sigma}_{\varepsilon}\right]^{-1}\left(\bar{\mathcal{Y}} - \mathbf{F}\widehat{\boldsymbol{\beta}}\right) + \mathbf{f}(\mathbf{x}_0)^\top\widehat{\boldsymbol{\beta}} \tag{38}$$

where $\widehat{\boldsymbol{\beta}}$ is the generalized least squares estimator (GLS estimator) of $\boldsymbol{\beta}$ corresponding to $\widehat{\mathsf{Y}}(\mathbf{x}_0)$:

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{F}^\top[\boldsymbol{\Sigma}_{\mathsf{M}} + \boldsymbol{\Sigma}_{\varepsilon}]^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^\top(\boldsymbol{\Sigma}_{\mathsf{M}} + \boldsymbol{\Sigma}_{\varepsilon})^{-1}\bar{\mathcal{Y}}\ . \tag{39}$$

The estimator $\widehat{\boldsymbol{\beta}}$ is also the best linear unbiased for $\boldsymbol{\beta}$. Finally by direct substitution, the MSE of the BLUP $\widehat{\mathsf{Y}}(\mathbf{x}_0)$ can be shown to be

$$\mathrm{MSE}^\star = \Sigma_{\mathsf{M}}(\mathbf{x}_0,\mathbf{x}_0) - \boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0,\cdot)^\top\left[\boldsymbol{\Sigma}_{\mathsf{M}} + \boldsymbol{\Sigma}_{\varepsilon}\right]^{-1}\boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0,\cdot) + \eta^\top(\mathbf{F}^\top[\boldsymbol{\Sigma}_{\mathsf{M}} + \boldsymbol{\Sigma}_{\varepsilon}]^{-1}\mathbf{F})^{-1}\eta. \tag{40}$$

where $\eta = \mathbf{f}(\mathbf{x}_0) - \mathbf{F}^\top\left[\boldsymbol{\Sigma}_{\mathsf{M}} + \boldsymbol{\Sigma}_{\varepsilon}\right]^{-1}\boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0,\cdot)$.

In the forthcoming sections, we use $\mathbf{f}(\mathbf{x}) = (1, x_1, x_2, \ldots, x_p)^\top$ to derive the results. In this case q equals p, the dimension of $\mathbf{x}$, hence $\mathbf{F}$ can be written as

$$\mathbf{F} = \begin{pmatrix} \mathbf{f}(\mathbf{x}_1)^\top \\ \mathbf{f}(\mathbf{x}_2)^\top \\ \vdots \\ \mathbf{f}(\mathbf{x}_k)^\top \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{k1} & \ldots & x_{kp} \end{pmatrix}.$$

A.2 Two-Point Intercept Model

For the k-point intercept model ($k \geq 2$), the BLUP $\widehat{\mathsf{Y}}(\mathbf{x}_0)$, $\widehat{\beta}_0$ and MSE* can be obtained by plugging $\mathbf{F} = \mathbf{1}_k$ and $\mathbf{f}(\mathbf{x}_0) = 1$ into Equations (38), (39) and (40). It follows that the BLUP of $\mathsf{Y}(\mathbf{x}_0)$ is

$$\widehat{\mathsf{Y}}(\mathbf{x}_0) = \widehat{\beta}_0 + \boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0,\cdot)^\top\left[\boldsymbol{\Sigma}_{\mathsf{M}} + \boldsymbol{\Sigma}_{\varepsilon}\right]^{-1}\left(\bar{\mathcal{Y}} - \widehat{\beta}_0\mathbf{1}_k\right) \tag{41}$$

where the corresponding BLUE of β_0 is

$$\widehat{\beta}_0 = (\mathbf{1}_k^\top\left[\boldsymbol{\Sigma}_{\mathsf{M}} + \boldsymbol{\Sigma}_{\varepsilon}\right]^{-1}\mathbf{1}_k)^{-1}\mathbf{1}_k^\top\left[\boldsymbol{\Sigma}_{\mathsf{M}} + \boldsymbol{\Sigma}_{\varepsilon}\right]^{-1}\bar{\mathcal{Y}} \tag{42}$$

with MSE

$$\mathrm{MSE}^\star = \Sigma_{\mathsf{M}}(\mathbf{x}_0,\mathbf{x}_0) - \boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0,\cdot)^\top\left[\boldsymbol{\Sigma}_{\mathsf{M}} + \boldsymbol{\Sigma}_{\varepsilon}\right]^{-1}\boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0,\cdot) + \eta^\top(\mathbf{1}_k^\top[\boldsymbol{\Sigma}_{\mathsf{M}} + \boldsymbol{\Sigma}_{\varepsilon}]^{-1}\mathbf{1}_k)^{-1}\eta \tag{43}$$

where $\eta = 1 - \mathbf{1}_k^\top\left[\boldsymbol{\Sigma}_{\mathsf{M}} + \boldsymbol{\Sigma}_{\varepsilon}\right]^{-1}\boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0,\cdot)$ and $\mathbf{1}_k$ denotes the $k \times 1$ vector of ones.

Compared to the result given in Equation (32) in Ankenman et al. (2010), it is worth noting that here an extra penalty term $\eta^\top(\mathbf{1}_k^\top[\boldsymbol{\Sigma}_{\mathsf{M}} + \boldsymbol{\Sigma}_{\varepsilon}]^{-1}\mathbf{1}_k)^{-1}\eta$ is incurred in MSE* for estimating the unknown parameter β_0.

Now for the two-point intercept model considered in Section 3.1, suppose that

$$\mathbf{\Sigma}_{\mathsf{M}} = \tau^2 \begin{pmatrix} 1 & r_{12} \\ r_{21} & 1 \end{pmatrix} \qquad \mathbf{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot) = \tau^2 \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} \quad \text{and} \quad \mathbf{\Sigma}_{\varepsilon} = \frac{\mathsf{V}}{n} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Using the Gaussian correlation function, we have $r_i = \exp\{-\theta(x_i - x_0)^2\}, i = 1, 2,$ and $r_{12} = \exp\{-\theta(x_1 - x_2)^2\}$. By plugging $\mathbf{\Sigma}_{\mathsf{M}}, \mathbf{\Sigma}_{\varepsilon}$ and $\mathbf{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot)$ into Equation (41), we obtain the expression for $\widehat{\mathsf{Y}}(x_0)$ shown in Equation (6). Then $\widehat{\nabla}_{\text{sk}}$ follows by taking the derivative of $\widehat{\mathsf{Y}}(x_0)$ with respect to x_0.

We next obtain the mean squared difference between the two gradient estimators obtained in the stochastic kriging setting when the number of replications n is finite and as $n \to \infty$. First observe the difference between the average responses across simulation replications at the design point x_i: As $n \to \infty$, $\bar{\mathcal{Y}}(x_i) = \beta_0 + \mathsf{M}(x_i)$; but when n is finite, $\bar{\mathcal{Y}}(x_i) = \beta_0 + \mathsf{M}(x_i) + \bar{\varepsilon}(x_i)$, where $i = 1, 2$ denotes the index of the design point.

It follows from Equation (9) that

$$\begin{aligned} \widehat{\nabla}_{\text{sk}}(n) - \widehat{\nabla}_{\text{sk}}(\infty) &= -\theta\tau^2 \left(r_1(x_0 - x_1) + r_2(x_2 - x_0)\right) \\ &\quad \times \left(\frac{[\mathsf{M}(x_1) + \bar{\varepsilon}(x_1)] - [\mathsf{M}(x_2) + \bar{\varepsilon}(x_2)]}{\tau^2(1 - r_{12}) + \frac{\mathsf{V}}{n}(1 - \rho)} - \frac{[\mathsf{M}(x_1) - \mathsf{M}(x_2)]}{\tau^2(1 - r_{12})} \right). \end{aligned}$$

The mean squared difference $\mathrm{E}\left[\widehat{\nabla}_{\text{sk}}(n) - \widehat{\nabla}_{\text{sk}}(\infty)\right]^2$ given in Equation (10) follows from some algebraic manipulations and recognizing the fact that $\mathsf{M}(\cdot)$ and $\bar{\varepsilon}(\cdot)$ are independent of each other.

A.3 k-Point Intercept Model

Since we have given the general expressions for the best linear predictor $\widehat{\mathsf{Y}}(\mathbf{x}_0)$, its corresponding $\widehat{\beta}_0$ and MSE* in Equations (41), (42) and (43) respectively, we will not repeat them here.

With respect to the k-point intercept model considered in Section 3.2, we have

$$\mathbf{\Sigma}_{\mathsf{M}} = \tau^2\, \mathbf{I}_k \qquad \text{and} \qquad \mathbf{\Sigma}_{\varepsilon} = \frac{\mathsf{V}}{n} \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}.$$

We also have $\mathbf{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot) = \tau^2(r_1, r_2, \ldots, r_k)^\top$. By plugging $\mathbf{\Sigma}_{\mathsf{M}}, \mathbf{\Sigma}_{\varepsilon}$ and $\mathbf{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot)$ into Equations (41) and (43), we get the expressions for $\widehat{\mathsf{Y}}(\mathbf{x}_0)$ and MSE* as shown in Equations (11) and (12).

Now since the ith point $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$ is a vector in $\Re^p, i = 0, 1, \ldots, k$, it follows that $\widehat{\nabla}_{\text{sk}}$ is a random vector in $\Re^p$. Let $\widehat{\nabla}_{\text{sk}} = \left(\widehat{\nabla}_{\text{sk}_1}, \widehat{\nabla}_{\text{sk}_2}, \ldots, \widehat{\nabla}_{\text{sk}_p}\right)^\top$. Taking the derivative of $\widehat{\mathsf{Y}}(\mathbf{x}_0)$ with respect to x_{0j} gives the jth component of this gradient

$$\begin{aligned}
\widehat{\nabla}_{\mathrm{sk}_j} &= \frac{\partial \widehat{\mathsf{Y}}(\mathbf{x}_0)}{\partial x_{0j}} \\
&= \frac{-2\theta\tau^2}{(\tau^2 + \frac{\mathsf{V}}{n}(1-\rho))} \cdot \sum_{i=1}^{k} \left(\left(\bar{\mathcal{Y}}(\mathbf{x}_i) - \frac{1}{k}\sum_{h=1}^{k} \bar{\mathcal{Y}}(\mathbf{x}_h) \right) (x_{0j} - x_{ij}) r_i \right) \qquad (44)
\end{aligned}$$

where $r_i = \exp\{-\theta \sum_{j=1}^{p}(x_{0j} - x_{ij})^2\}$ is the spatial correlation between $\mathbf{x}_i$ and $\mathbf{x}_0$. Notice that $\widehat{\nabla}_{\mathrm{sk}_j}(\infty)$ can be obtained by substituting $\mathsf{V} = 0$ into Equation (44).

Therefore, the difference between the jth gradient components is

$$\begin{aligned}
\widehat{\nabla}_{\mathrm{sk}_j}(n) - \widehat{\nabla}_{\mathrm{sk}_j}(\infty) &= \frac{-2\theta\tau^2}{(\tau^2 + \frac{\mathsf{V}}{n}(1-\rho))} \sum_{i=1}^{k} \left(a_{ij} \left(\mathsf{M}(\mathbf{x}_i) + \bar{\varepsilon}(\mathbf{x}_i) - \frac{1}{k}\sum_{h=1}^{k}(\mathsf{M}(\mathbf{x}_h) + \bar{\varepsilon}(\mathbf{x}_h)) \right) \right) \\
&\quad -(-2\theta) \sum_{i=1}^{k} \left(a_{ij} \left(\mathsf{M}(\mathbf{x}_i) - \frac{1}{k}\sum_{h=1}^{k} \mathsf{M}(\mathbf{x}_h) \right) \right)
\end{aligned}$$

and the mean square of this difference is

$$\mathrm{E}\left[\left(\widehat{\nabla}_{\mathrm{sk}_j}(n) - \widehat{\nabla}_{\mathrm{sk}_j}(\infty) \right)^2 \right] = \frac{4\theta^2}{\left(\frac{1}{\frac{\mathsf{V}}{n}(1-\rho)} + \frac{1}{\tau^2} \right)} \left(\sum_{i=1}^{k} a_{ij}^2 - \frac{1}{k} (\sum_{i=1}^{k} a_{ij})^2 \right)$$

where $a_{ij} = (x_{0j} - x_{ij})\, r_i$, $i = 1, 2, \ldots, k; j = 1, 2, \ldots, p$.

We evaluate the "distance" between the two gradient estimators by the inner product of the difference between the two random vectors $\widehat{\nabla}_{\mathrm{sk}}(n)$ and $\widehat{\nabla}_{\mathrm{sk}}(\infty)$ at prediction point $\mathbf{x}_0 \in \Re^p$:

$$\begin{aligned}
\langle \widehat{\nabla}_{\mathrm{sk}}(n) - \widehat{\nabla}_{\mathrm{sk}}(\infty), \widehat{\nabla}_{\mathrm{sk}}(n) - \widehat{\nabla}_{\mathrm{sk}}(\infty) \rangle &= \sum_{j=1}^{p} \mathrm{E}\left[\left(\widehat{\nabla}_{\mathrm{sk}_j}(n) - \widehat{\nabla}_{\mathrm{sk}_j}(\infty) \right)^2 \right] \\
= \frac{4\theta^2}{\left(\frac{1}{\frac{\mathsf{V}}{n}(1-\rho)} + \frac{1}{\tau^2} \right)} \sum_{j=1}^{p} \left(\sum_{i=1}^{k} a_{ij}^2 - \frac{1}{k} (\sum_{i=1}^{k} a_{ij})^2 \right) \\
= \frac{4\theta^2}{\left(\frac{1}{\frac{\mathsf{V}}{n}(1-\rho)} + \frac{1}{\tau^2} \right)} \sum_{j=1}^{p} \left(\sum_{i=1}^{k} (x_{0j} - x_{ij})^2 r_i^2 - \frac{1}{k} \left(\sum_{i=1}^{k} (x_{0j} - x_{ij}) r_i \right)^2 \right).
\end{aligned}$$

A.4 Two-Point Trend model

By the results in Appendix A.1, for the two-point trend model ($k = 2$), the BLUP $\widehat{\mathsf{Y}}(x_0)$, its corresponding $\widehat{\beta}_0, \widehat{\beta}_1$ and MSE* can be obtained by plugging $\mathbf{F} = \mathbf{D}_2$ and $\mathbf{f}(x_0) = (1, x_0)^\top$

into Equations (38), (39) and (40). The 2×2 design matrix $\mathbf{D}_2$ is as specified in Section 4.1

$$\mathbf{D}_2 = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \end{pmatrix}.$$

Simplification of Equation (39) gives us $\widehat{\boldsymbol{\beta}} = \mathbf{D}_2^{-1}\bar{\mathcal{Y}}$, and the unbiasedness of $\widehat{\boldsymbol{\beta}}$ follows immediately from $\mathrm{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{D}_2^{-1}\mathrm{E}(\bar{\mathcal{Y}}) = \mathbf{D}_2^{-1}\mathbf{D}_2\boldsymbol{\beta} = \boldsymbol{\beta}$. The variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ is $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = (\mathbf{D}_2^\top[\boldsymbol{\Sigma}_{\mathsf{M}} + \boldsymbol{\Sigma}_\varepsilon]^{-1}\mathbf{D}_2)^{-1}$ and this gives $\mathrm{Var}(\beta_0)$, $\mathrm{Var}(\beta_1)$ and $\mathrm{Cov}(\beta_0, \beta_1)$. Notice that in this case $\widehat{\mathsf{Y}}(\mathbf{x}_0)$ can be reduced into a simple form

$$\widehat{\mathsf{Y}}(x_0) = \boldsymbol{\Sigma}_{\mathsf{M}}(x_0,\cdot)^\top\left[\boldsymbol{\Sigma}_{\mathsf{M}} + \boldsymbol{\Sigma}_\varepsilon\right]^{-1}\left(\bar{\mathcal{Y}} - \mathbf{D}_2\mathbf{D}_2^{-1}\bar{\mathcal{Y}}\right) + \mathbf{f}(x_0)^\top\widehat{\boldsymbol{\beta}} = \mathbf{f}(x_0)^\top\mathbf{D}_2^{-1}\bar{\mathcal{Y}}$$

as shown in Equation (16). Finally, the gradient $\widehat{\nabla}_{\text{sk}}$ of $\widehat{\mathsf{Y}}(x_0)$ is obtained by taking the derivative of $\widehat{\mathsf{Y}}(x_0)$ with respect to x_0

$$\widehat{\nabla}_{\text{sk}} = \frac{d\widehat{\mathsf{Y}}(x_0)}{dx_0} = \frac{d(\mathbf{f}(x_0))^\top}{dx_0}\mathbf{D}_2^{-1}\bar{\mathcal{Y}} = \widehat{\beta}_1 = \frac{\bar{\mathcal{Y}}(x_2) - \bar{\mathcal{Y}}(x_1)}{x_2 - x_1}\ .$$

We next obtain the mean squared difference between the two gradient estimators obtained in the stochastic kriging setting when the number of replications n is finite and as $n\to\infty$.

$$\begin{aligned} \mathrm{E}\left[\widehat{\nabla}_{\text{sk}}(n) - \widehat{\nabla}_{\text{sk}}(\infty)\right]^2 &= \frac{1}{(x_2-x_1)^2}\mathrm{E}\left[\bar{\varepsilon}(x_2) - \bar{\varepsilon}(x_1)\right]^2 \\ &= \frac{2\mathsf{V}(1-\rho)}{n(x_1-x_2)^2}\ . \end{aligned}$$

A.5 k-Point Trend Model

For the results proved below for the k-point trend model, we use the same $\boldsymbol{\Sigma}_{\mathsf{M}}$ and $\boldsymbol{\Sigma}_\varepsilon$ as for the k-point intercept model in Section 3.2; however, we assume that all the design points have identical spatial correlation to the prediction point $\mathbf{x}_0$, that is, $\boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0,\cdot) = \tau^2(r_0, r_0, \ldots, r_0)^\top$.

For the k-point trend model ($k > 2$), by the results in Appendix A.1, the BLUP $\widehat{\mathsf{Y}}(x_0)$, the corresponding BLUE $\widehat{\boldsymbol{\beta}}$ and MSE* can be obtained by substituting $\mathbf{F} = \mathbf{D}_k$ and $\mathbf{f}(\mathbf{x}_0) = (1, x_{01}, x_{02}, \ldots, x_{0p})^\top$ into Equations (38), (39) and (40). The $k\times(p+1)$ orthogonal design matrix $\mathbf{D}_k$ of rank $(p+1)$ is as specified in Section 4.2:

$$\mathbf{D}_k = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{k1} & \ldots & x_{kp} \end{pmatrix}$$

where the column vectors of $\mathbf{D}_k$ are pairwise orthogonal. Equation (39) gives us $\widehat{\boldsymbol{\beta}} = (\mathbf{D}_k^\top[\boldsymbol{\Sigma}_{\mathsf{M}} + \boldsymbol{\Sigma}_\varepsilon]^{-1}\mathbf{D}_k)^{-1}\mathbf{D}_k^\top[\boldsymbol{\Sigma}_{\mathsf{M}}+\boldsymbol{\Sigma}_\varepsilon]^{-1}\bar{\mathcal{Y}}$; and the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ is $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) =$

$(\mathbf{D}_k^\top[\mathbf{\Sigma}_{\mathsf{M}}+\mathbf{\Sigma}_\varepsilon]^{-1}\mathbf{D}_k)^{-1}$. Notice that the orthogonality assumption enables us to simplify these expressions to a great extent, as shown in Equations (23), (24) and (28).

For $j = 1, 2, \ldots, p$, the jth component of the gradient is

$$\begin{aligned}\widehat{\nabla}_{\mathrm{sk}_j} &= \frac{\partial \widehat{\mathsf{Y}}(\mathbf{x}_0)}{\partial x_{0j}} \\ &= \frac{d\mathbf{\Sigma}_{\mathsf{M}}(\mathbf{x}_0,\cdot)}{dx_{0j}}[\mathbf{\Sigma}_{\mathsf{M}}+\mathbf{\Sigma}_\varepsilon]^{-1}(\bar{\mathcal{Y}}-\mathbf{D}_2\widehat{\boldsymbol{\beta}})+\widehat{\beta}_j \\ &= \widehat{\beta}_j = \frac{\sum_{i=1}^k x_{ij}\bar{\mathcal{Y}}(\mathbf{x}_i)}{\sum_{i=1}^k x_{ij}^2}\quad .\end{aligned}$$

Hence the difference between the jth gradient components is

$$\begin{aligned}\widehat{\nabla}_{\mathrm{sk}_j}(n)-\widehat{\nabla}_{\mathrm{sk}_j}(\infty) &= \frac{\sum_{i=1}^k x_{ij}\left(\beta_0+\sum_{d=1}^p\beta_d x_{id}+\mathsf{M}(\mathbf{x}_i)+\bar{\varepsilon}(\mathbf{x}_i)\right)}{\sum_{i=1}^k x_{ij}^2} \\ &\quad -\frac{\sum_{i=1}^k x_{ij}\left(\beta_0+\sum_{d=1}^p\beta_d x_{id}+\mathsf{M}(\mathbf{x}_i)\right)}{\sum_{i=1}^k x_{ij}^2} = \frac{\sum_{i=1}^k x_{ij}\bar{\varepsilon}(\mathbf{x}_i)}{\sum_{i=1}^k x_{ij}^2}\end{aligned}$$

and the mean square of this difference is

$$\begin{aligned}\mathrm{E}\left[\left(\widehat{\nabla}_{\mathrm{sk}_j}(n)-\widehat{\nabla}_{\mathrm{krig}_j}(\infty)\right)^2\right] &= \frac{\mathrm{E}\left[\left(\sum_{i=1}^k x_{ij}\bar{\varepsilon}(\mathbf{x}_i)\right)^2\right]}{\left(\sum_{i=1}^k x_{ij}^2\right)^2} \\ &= \frac{\mathrm{Var}(\mathbf{x}_{\cdot j}^\top\cdot\bar{\varepsilon})}{\left(\sum_{i=1}^k x_{ij}^2\right)^2} = \frac{\mathbf{x}_{\cdot j}^\top\mathbf{\Sigma}_\varepsilon\mathbf{x}_{\cdot j}}{\left(\sum_{i=1}^k x_{ij}^2\right)^2} \\ &= \frac{\mathsf{V}(1-\rho)}{n\left(\sum_{i=1}^k x_{ij}^2\right)}\end{aligned}$$

where $\mathbf{x}_{\cdot j} = (x_{1j}, x_{2j}, \ldots, x_{kj})^\top, j = 1, 2, \ldots, p$; $\bar{\varepsilon} = (\bar{\varepsilon}(\mathbf{x}_1), \bar{\varepsilon}(\mathbf{x}_2), \ldots, \bar{\varepsilon}(\mathbf{x}_k))^\top$. The last step is a consequence of the orthogonality assumption which implies that $\sum_{i=1}^k x_{ij} = 0$ in a design that contains a column of 1's.

As for the k-point intercept model, the mean squared difference is defined as the inner product of the difference between the two random vectors $\widehat{\nabla}_{\mathrm{sk}}(n)$ and $\widehat{\nabla}_{\mathrm{sk}}(\infty)$ at prediction point $\mathbf{x}_0 \in \Re^p$:

$$\begin{aligned}\langle\widehat{\nabla}_{\mathrm{sk}}(n)-\widehat{\nabla}_{\mathrm{sk}}(\infty),\widehat{\nabla}_{\mathrm{sk}}(n)-\widehat{\nabla}_{\mathrm{sk}}(\infty)\rangle &= \sum_{j=1}^p\mathrm{E}\left[\left(\widehat{\nabla}_{\mathrm{sk}_j}(n)-\widehat{\nabla}_{\mathrm{sk}_j}(\infty)\right)^2\right] \\ &= \frac{\mathsf{V}(1-\rho)}{n}\sum_{j=1}^p\left(\sum_{i=1}^k x_{ij}^2\right)^{-1}.\end{aligned}$$

A.6 Two-point Intercept Model: MSE Derivative

To see the effect of CRN on this optimal MSE*, we take the derivative of Equation (7) with respect to ρ:

$$\frac{d\text{MSE}^\star}{d\rho} = \frac{1}{2}\frac{\mathsf{V}}{n}\left(1 - \frac{(r_1 - r_2)^2}{[(1 - r_{12}) + (1 - \rho)\gamma/n]^2}\right) \tag{45}$$

where $\gamma = \mathsf{V}/\tau^2$ is defined as the ratio of the intrinsic variance to the extrinsic variance. We claim that $d\text{MSE}^\star/d\rho \geq 0$. Consider a fixed experimental design, where x_1 and x_2 and hence r_{12} have been fixed. Other things being equal (including ρ), the sign of $d\text{MSE}^\star/d\rho$ is determined by the squared difference between r_1 and r_2. Now when x_1 and x_2 have identical spatial correlation with x_0, we see that $d\text{MSE}^\star/d\rho > 0$, hence CRN will increase MSE*. This is the result proved in Ankenman et al. (2010). We know that the greater the distance between two points (in any appropriate norm), the smaller the spatial correlation between them. Now we consider the case that $r_1 \neq r_2$. Without loss of generality, suppose that $r_1 > r_2$. If the prediction point x_0 is located between x_1 and x_2, then the prediction point x_0 should be closer to x_1 than to x_2. As we move x_0 even closer to x_1, $(r_1 - r_2)$ increases. As x_0 gets very close to x_1, $(r_1 - r_2)$ gets very close to $(1 - r_{12})$. In this extreme case, $d\text{MSE}^\star/d\rho$ is still positive because of the non-negative term $(1 - \rho)\gamma/n$ in the denominator. Hence as long as the intrinsic correlation $\rho \geq 0$, it is always true that $d\text{MSE}^\star/d\rho \geq 0$, that is, CRN increases MSE*.

A.7 k-Point Intercept Model: MSE Derivative

To see the effect of CRN on MSE*, we take its derivative of Equation (12) with respect to ρ:

$$\frac{d\text{MSE}^\star}{d\rho} = \frac{1}{k}\frac{\mathsf{V}}{n}\left((k-1) - \frac{k\sum_{i=1}^k r_i^2 - (\sum_{i=1}^k r_i)^2}{(1 + (1 - \rho)\gamma/n)^2}\right). \tag{46}$$

The proof is easy to follow once we notice that the condition for the spatial variance-covariance matrix of $(\mathsf{Y}(\mathbf{x}_0), \bar{\mathcal{Y}}(\mathbf{x}_1), \ldots, \bar{\mathcal{Y}}(\mathbf{x}_k))^\top$ to be positive definite is $\sum_{i=1}^k r_i^2 < 1$. By rewriting the numerator in the second term in Equation (46), we get

$$k\sum_{i=1}^k r_i^2 - \left(\sum_{i=1}^k r_i\right)^2 = (k-1)\sum_{i=1}^k r_i^2 + \left(\sum_{i=1}^k {r_i}^2 - \left(\sum_{i=1}^k r_i\right)^2\right) < k - 1 \quad .$$

Furthermore, the denominator $(1+(1-\rho)\gamma/n)^2$ is always greater than or equal to 1. Therefore $d\text{MSE}^\star/d\rho$ is positive for any $\rho \in [0, 1)$, and this implies that CRN increases MSE*.
Finally, observe that by substituting $k = 2$ in Equation (46), we get the same conclusion that was drawn from Equation (45) for the two-point intercept problem when the two design points are assumed to be spatially uncorrelated.

A.8 Proof of Theorem 1

The proof follows the same line of argument as in Ankenman et al. (2010). First we show that for any fixed positive definite covariance matrix $\boldsymbol{\Sigma}'_{\varepsilon}$ and $k \times (q+1)$ design matrix $\mathbf{F}$ of full rank, the predictor

$$\widehat{\mathsf{Y}}'(\mathbf{x}_0) = \mathbf{f}(\mathbf{x}_0)^{\top}\boldsymbol{\beta} + \boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0,\cdot)^{\top}\left[\boldsymbol{\Sigma}_{\mathsf{M}} + \boldsymbol{\Sigma}'_{\varepsilon}\right]^{-1}\left(\bar{\mathcal{Y}} - \mathbf{F}\boldsymbol{\beta}\right)$$

is an unbiased predictor. This follows immediately from $\mathrm{E}[\widehat{\mathsf{Y}}'(\mathbf{x}_0) - \mathsf{Y}(\mathbf{x}_0)] = \mathbf{f}(\mathbf{x}_0)^{\top}\boldsymbol{\beta} + 0 - \mathbf{f}(\mathbf{x}_0)^{\top}\boldsymbol{\beta} = 0$.

Next notice that under Model (2), we have

$$\begin{aligned}\mathcal{S}_{i\ell} &= \frac{1}{n-1}\sum_{j=1}^{n}(\mathcal{Y}_j(\mathbf{x}_i) - \bar{\mathcal{Y}}(\mathbf{x}_i))(\mathcal{Y}_j(\mathbf{x}_\ell) - \bar{\mathcal{Y}}(\mathbf{x}_\ell)) \\ &= \frac{1}{n-1}\sum_{j=1}^{n}(\varepsilon_j(\mathbf{x}_i) - \bar{\varepsilon}(\mathbf{x}_i))(\varepsilon_j(\mathbf{x}_\ell) - \bar{\varepsilon}(\mathbf{x}_\ell)).\end{aligned}$$

where $\mathcal{S}_{i\ell}$ is the sample covariance of the random noise at design points $\mathbf{x}_i$ and $\mathbf{x}_\ell$, $i, \ell = 1, 2, \ldots, k$. We use $\widehat{\boldsymbol{\Sigma}}_{\varepsilon} = n^{-1}\mathcal{S}$ to estimate $\boldsymbol{\Sigma}_{\varepsilon}$. Therefore, under Assumption 1, $\widehat{\mathsf{V}}(\mathbf{x}_i) = \mathcal{S}_{ii}$ is independent of $\bar{\mathcal{Y}}(\mathbf{x}_i) = \mathbf{f}(\mathbf{x}_i)^{\top}\boldsymbol{\beta} + \mathsf{M}(\mathbf{x}_i) + \bar{\varepsilon}(\mathbf{x}_i)$ by the properties of the multivariate normal distribution (recall that M is also independent of $\varepsilon_j(\mathbf{x}_i)$). Then since $\widehat{\boldsymbol{\Sigma}}_{\varepsilon} = n^{-1}\mathcal{S}$, it follows that

$$\begin{aligned}\mathrm{E}\left[\widehat{\widehat{\mathsf{Y}}}(\mathbf{x}_0) - \mathsf{Y}(\mathbf{x}_0)\right] &= \mathrm{E}\left[\mathrm{E}\left(\widehat{\widehat{\mathsf{Y}}}(\mathbf{x}_0) - \mathsf{Y}(\mathbf{x}_0)\,\middle|\,\widehat{\boldsymbol{\Sigma}}_{\varepsilon}\right)\right] \\ &= \mathrm{E}\left[\mathbf{f}(\mathbf{x}_0)^{\top}\boldsymbol{\beta} - \mathbf{f}(\mathbf{x}_0)^{\top}\boldsymbol{\beta}\right] = 0.\end{aligned}$$

A.9 MSE Inflation of Section 5

Here we prove the results in Section 5 on MSE inflation, when every parameter is known except the intrinsic variance V. Suppose that we have

$$\boldsymbol{\Sigma}_{\mathsf{M}} = \tau^2\begin{pmatrix} 1 & r & \cdots & r \\ r & 1 & \cdots & r \\ \vdots & \vdots & \ddots & \vdots \\ r & r & \cdots & 1 \end{pmatrix} \qquad \boldsymbol{\Sigma}_{\varepsilon} = \frac{\mathsf{V}}{n}\begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

and $\boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0,\cdot) = \tau^2(r_0, r_0, \ldots, r_0)^{\top}$ with $r_0, r \geq 0$ and we have an independent estimator $\widehat{\mathsf{V}} \sim \mathsf{V}\chi^2_{n-1}/(n-1)$. Let $\gamma = \mathsf{V}/\tau^2$ be the ratio of the intrinsic variance to the extrinsic variance. A key to the result is noting that we can write

$$\boldsymbol{\Sigma}_{\mathsf{M}} + \boldsymbol{\Sigma}_{\varepsilon} = \mathbf{q}\mathbf{q}^T + \left(\tau^2(1-r) + \frac{\mathsf{V}}{n}(1-\rho)\right)\cdot\mathbf{I}_k$$

where $\mathbf{q}^T = \sqrt{\tau^2 r + (\mathsf{V}/n)\rho} \cdot \mathbf{1}_k^\top$, $\mathbf{I}_k$ is the $k \times k$ identity matrix, and $\mathbf{1}_k$ is the $k \times 1$ vector of ones. This matrix is invertible in closed form using Theorem 8.3.3 of Graybill (1969). Specialized to this case

$$\begin{aligned}
[\boldsymbol{\Sigma}_\mathsf{M} + \boldsymbol{\Sigma}_\varepsilon]^{-1} &= \left(\frac{1}{\tau^2(1-r) + \frac{\mathsf{V}}{n}(1-\rho)}\right) \cdot \mathbf{I}_k \\
&\quad - \frac{\tau^2 r + \frac{\mathsf{V}}{n}\rho}{\left(\tau^2 r + \frac{\mathsf{V}}{n}\rho\right)(k-1) + \left(\tau^2 + \frac{\mathsf{V}}{n}\right)} \left\| \frac{1}{\tau^2(1-r) + \frac{\mathsf{V}}{n}(1-\rho)} \right\| \qquad (47)
\end{aligned}$$

where $\|\cdot\|$ indicates a matrix of appropriate dimension with all elements the same.

We obtain the MSE of $\widehat{\mathsf{Y}}(\mathbf{x}_0)$, the stochastic kriging BLUP with $\boldsymbol{\Sigma}_\varepsilon$ (i.e., V) known,

$$\mathrm{MSE}^\star = \tau^2 \left(1 - \frac{k r_0^2}{1 + C_{\rho\gamma} + (k-1)r}\right)$$

by plugging $\boldsymbol{\Sigma}_\mathsf{M}, \boldsymbol{\Sigma}_\varepsilon$ and $\boldsymbol{\Sigma}_\mathsf{M}(\mathbf{x}_0, \cdot)$ into (43) and using the known inverse (47).

The MSE of $\widehat{\widehat{\mathsf{Y}}}(\mathbf{x}_0)$ is derived by substituting $\widehat{\mathsf{V}}$ for V and noting that since $\widehat{\widehat{\mathsf{Y}}}(\mathbf{x}_0)$ is unbiased for any $\widehat{\mathsf{V}}$ independent of $\bar{\mathcal{Y}}$ (see Theorem 1 and the proof in Section A.8)

$$\begin{aligned}
\mathrm{MSE}\left[\widehat{\widehat{\mathsf{Y}}}(\mathbf{x}_0)\right] &= \mathrm{Var}\left[\widehat{\widehat{\mathsf{Y}}}(\mathbf{x}_0)\right] \\
&= \mathrm{E}\left[\mathrm{Var}\left(\widehat{\widehat{\mathsf{Y}}}(\mathbf{x}_0) \middle| \widehat{\mathsf{V}}\right)\right] + \mathrm{Var}\left[\mathrm{E}\left(\widehat{\widehat{\mathsf{Y}}}(\mathbf{x}_0) \middle| \widehat{\mathsf{V}}\right)\right] \\
&= \mathrm{E}\left[\mathrm{Var}\left(\widehat{\widehat{\mathsf{Y}}}(\mathbf{x}_0) \middle| \widehat{\mathsf{V}}\right)\right] \\
&= \mathrm{E}\left[\mathrm{Var}\left(\beta_0 + \tau^2 (r_0, r_0, \ldots, r_0)^\top \left[\boldsymbol{\Sigma}_\mathsf{M} + \frac{\widehat{\mathsf{V}}}{\mathsf{V}} \boldsymbol{\Sigma}_\varepsilon\right]^{-1} \left(\bar{\mathcal{Y}} - \beta_0 \mathbf{1}_k\right) \middle| \widehat{\mathsf{V}}\right)\right] \\
&= \tau^2 \mathrm{E}\left[1 + \frac{k r_0^2 \left[1 + C_{\rho\gamma} + (k-1)r\right]}{\left(1 + \frac{\widehat{\mathsf{V}}}{\mathsf{V}} C_{\rho\gamma} + (k-1)r\right)^2} - \frac{2 k r_0^2}{\left(1 + \frac{\widehat{\mathsf{V}}}{\mathsf{V}} C_{\rho\gamma} + (k-1)r\right)}\right]
\end{aligned}$$

where $C_{\rho\gamma} = \frac{\gamma}{n}(1 + (k-1)\rho)$ and $\gamma = \mathsf{V}/\tau^2$. The last step requires writing out the conditional variance, recalling that $\bar{\mathcal{Y}}$ and $\widehat{\mathsf{V}}$ are independent, and employing several tedious applications of Theorem 8.3.3 of Graybill (1969).

A.10 The Role of CRN on MSE Inflation of Section 5

Equations (33) and (34) in Section 5 suggest an enlightening way of doing this numerical analysis efficiently. Observe that in both equations, any combination of γ and ρ that gives

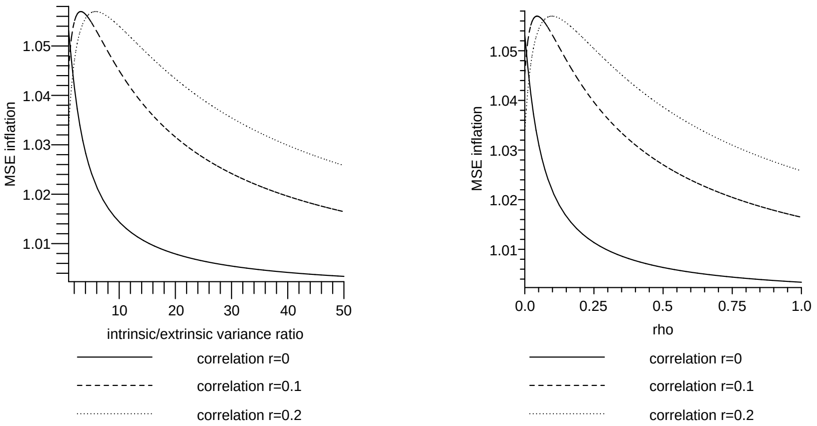

Figure 5: Illustration of equivalent impacts on MSE inflation by varying ρ from 0 to 1 while fixing $\gamma = 40$ and by varying γ from 1 to 50 while fixing $\rho = 39/49$, when $n = 10$ and correlation r_0 is 95% of its maximum possible value. On the left: $\rho = 39/49$ and $\gamma \in [1, 50]$; on the right: $\gamma = 40$ and $\rho \in [0, 1]$.

the same $C_{\rho\gamma}$ gives the same MSE inflation ratio. For instance, for any fixed γ_0 in $[1, k]$, we have

$$C_{\rho\gamma}(\rho_0, \gamma_0) = \frac{\gamma_0}{n}(1 + (k-1)\rho_0) = C_{\rho\gamma}(\rho', \gamma') \tag{48}$$

where $\rho' = (\gamma_0 - 1)/(k-1)$ and $\gamma' = 1 + (k-1)\rho_0$. Equation (48) reveals the symmetric roles played by ρ and γ on the MSE inflation: for a fixed $\gamma \in [1, k]$, the effect of CRN on the MSE inflation ratio by varying ρ from 0 to 1, is equivalent to that of fixing $\rho = (\gamma - 1)/(k-1)$ and varying γ from 1 to k. We illustrate this idea in Figure 5.

We consider the following three intervals for γ: $[0, 1], (1, k]$ and $(k, 10k]$, where $k = 50$ is the number of design points used in the numerical analysis.

- $\gamma = \mathsf{V}/\tau^2 \in [0, 1]$: As shown in Figure 6, when the intrinsic variance V is almost as large as the spatial variance τ^2, CRN increases the MSE inflation ratio. However, the stronger the spatial correlation among design points, the less that CRN inflates the MSE.

- $\gamma = \mathsf{V}/\tau^2 \in (1, 50]$: As shown in Figure 7, the majority of each curve is transitting from concave to convex, and the peak is shifting to the left, toward the direction of smaller ρ. Notice that the ordering of MSE inflation ratio according to the magnitude of the spatial correlation is reversing gradually as γ gets larger. In other words, when the intrinsic variance is larger than the spatial variance, CRN helps reduce the MSE inflation ratio when ρ is not very small. The less spatial correlation among design points, the more CRN reduces the MSE inflation.

- $\gamma = \mathsf{V}/\tau^2 \in (50, 500]$: As shown in Figure 8, the peak of each curve shifts to the very left edge, therefore each curve looks mostly monotone and convex. We see that when

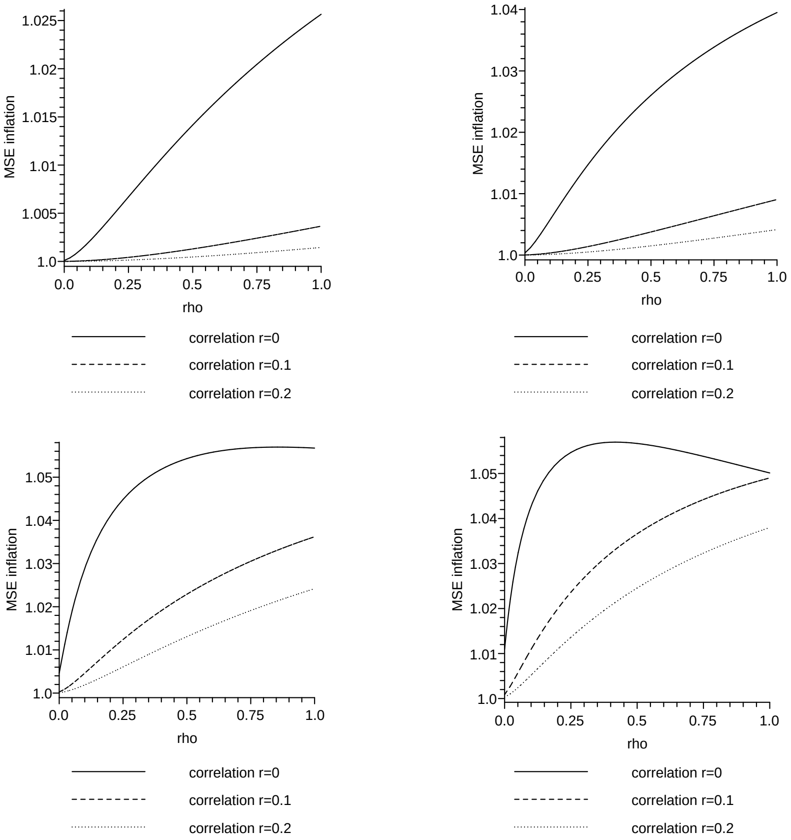

Figure 6: MSE inflation as a function of ρ when $n = 10$ and correlation r_0 is 95% of its maximum possible value. Top: from left to right: $\gamma = 0.05$ and 0.1; bottom: $\gamma = 0.5$ and 1.

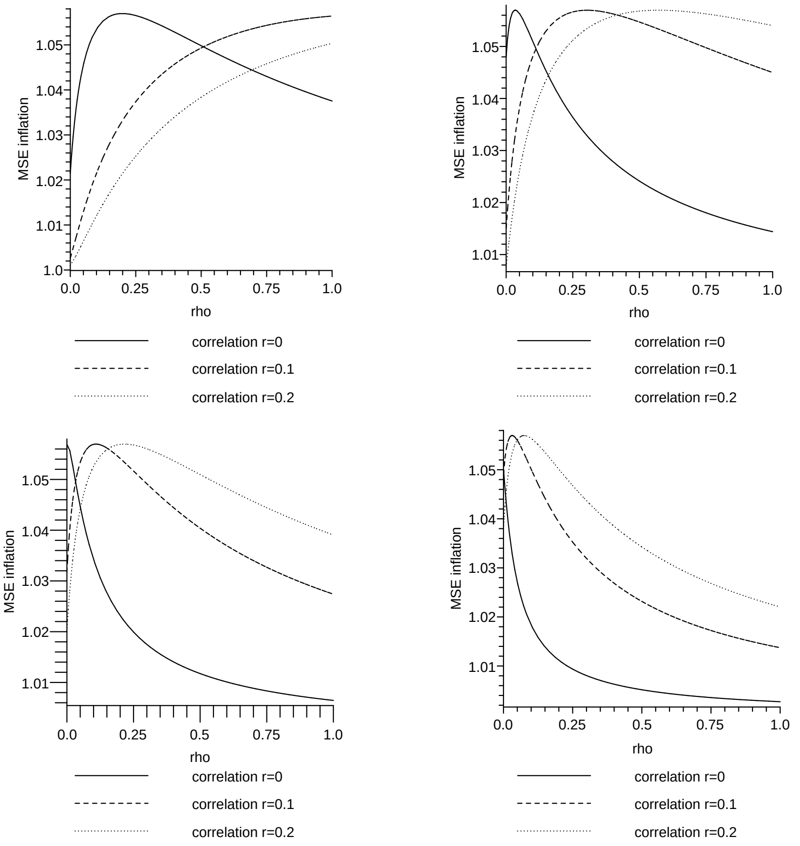

Figure 7: MSE inflation as a function of ρ when $n = 10$ and correlation r_0 is 95% of its maximum possible value. Top: from left to right: $\gamma = 2$ and 8; bottom: $\gamma = 20$ and 50.

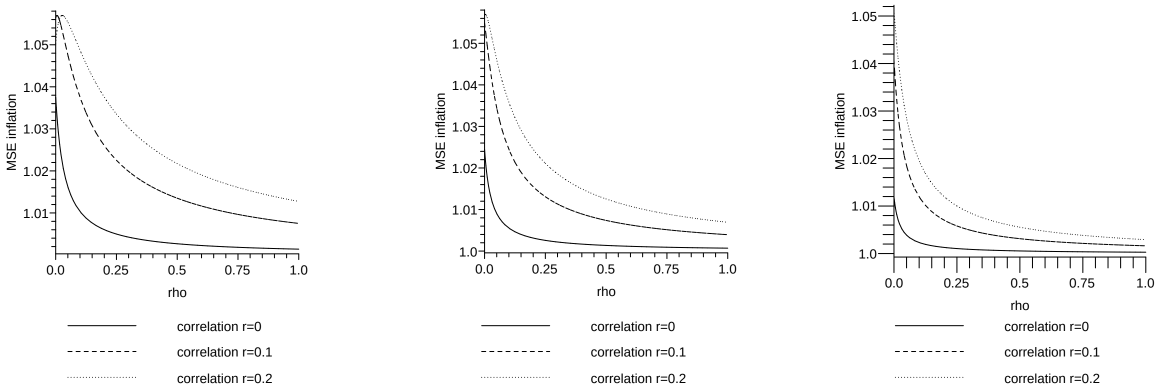

Figure 8: MSE inflation as a function of ρ when $n = 10$ and correlation r_0 is 95% of its maximum possible value. From left to right: $\gamma = 100, 200$, and 500.

the intrinsic variance dominates the spatial variance considerably, CRN helps reduce the MSE inflation ratio as long as the induced intrinsic correlation ρ is not very small. Furthermore, the smaller the spatial correlation among design points, the more CRN reduces the MSE inflation ratio.

References

Graybill, F. A. 1969. *Matrices with Applications in Statistics*, 2nd edition, Wadsworth, Belmont, CA.

Stein, M. L. 1999. *Interpolation of Spatial Data: Some Theory for Kriging.* Springer, NY.