

# Particle Swarm Optimization for Spatial Design

Matthew Simpson<sup>1</sup>

## Abstract

## KEY WORDS:

---

<sup>1</sup>(to whom correspondence should be addressed) Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO 65211-6100, themattsimpson@gmail.com

# 1 The Problem

Suppose we are interested in the latent spatial field of some response variable  $Y(\mathbf{s})$ ,  $\mathbf{s} \in \mathcal{D} \subseteq \mathbb{R}^2$ . Specifically, we are interested in predicting  $Y(\mathbf{s})$  at a set of target locations  $\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_M \in \mathcal{D}$ . We have the ability to sample  $N$  locations anywhere in  $\mathcal{D}$ , and we wish to place them in order to optimize some design criterion. Let  $\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_N \in \mathcal{D}$  denote the  $N$  sampled locations. Suppose that  $Y(\mathbf{s})$  is a geostatistical process with mean function  $\mu(\mathbf{s}) = \mathbf{x}(\mathbf{s})'\boldsymbol{\beta}$  for some known covariate  $\mathbf{x}(\mathbf{s})$  and covariance function  $C(\mathbf{s}, \mathbf{t})$  for  $\mathbf{s}, \mathbf{t} \in \mathcal{D}$ . Once the design points are selected, we observe  $Z(\mathbf{d}_i)$  for  $i = 1, 2, \dots, N$  where  $Z(\mathbf{d}) = Y(\mathbf{d}) + \varepsilon(\mathbf{d})$  and  $\varepsilon(\mathbf{d})$  is mean zero white noise with variance  $\sigma_\varepsilon^2$ , representing measurement error. An intuitive criterion to minimize is the average mean square prediction error (MSPE) from kriging across each of the target locations, each the errors are weighted appropriately. It turns out that this is intimately related to a more principled design criterion, the expected entropy gain on the predictive distribution:  $E\{\log[Y(\mathbf{s}_0)|\mathbf{Z}] - \log[Y(\mathbf{s}_0)]\}$ . Since the prior does not depend on the design, this is equivalent to maximizing  $E\{\log[Y(\mathbf{s}_0)|\mathbf{Z}]\}$ .

## 1.1 Simple Kriging

In simple kriging,  $C(\cdot, \cdot)$ ,  $\boldsymbol{\beta}$ , and  $\sigma_\varepsilon^2$  are all treated as known. Let  $\mathbf{Z} = [Z(\mathbf{d}_1), Z(\mathbf{d}_2), \dots, Z(\mathbf{d}_N)]'$ ,  $\mathbf{X} = [\mathbf{x}(\mathbf{d}_1), \mathbf{x}(\mathbf{d}_2), \dots, \mathbf{x}(\mathbf{d}_N)]'$ ,  $\mathbf{C}_Z = \text{cov}(\mathbf{Z})$  where  $\text{cov}[Z(\mathbf{d}_i), Z(\mathbf{d}_j)] = C(\mathbf{d}_i, \mathbf{d}_j) + \sigma_\varepsilon^2 1(\mathbf{d}_i = \mathbf{d}_j)$ , and  $\mathbf{c}_Y(\mathbf{s}_0) = \text{cov}[Y(\mathbf{s}_0), \mathbf{Z}]$  where  $\text{cov}[Y(\mathbf{s}_0), Z(\mathbf{d}_i)] = C(\mathbf{s}_0, \mathbf{d}_i)$ . Then the simple kriging predictor of  $Y(\mathbf{s}_0)$  is the linear predictor  $\mathbf{Y}^*(\mathbf{s}_0) = \mathbf{X}'\mathbf{Z} + k$  that MSPE conditional on sampled locations,  $\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_N$ . Specifically it minimizes  $E[Y(\mathbf{s}_0) - \hat{Y}_{sk}(\mathbf{s}_0)]^2$  over  $\boldsymbol{\lambda}$  and  $k$  such that  $\mathbf{X}'\mathbf{1} = 1$  where  $\mathbf{1}$  is a column vector of ones. The simple kriging predictor is easily derived as

$$\hat{Y}_{sk}(\mathbf{s}_0) = \mathbf{x}(\mathbf{s}_0)'\boldsymbol{\beta} + \mathbf{c}_Y(\mathbf{s}_0)'\mathbf{C}_Z^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})$$

with MSPE

$$\sigma_{sk}^2(\mathbf{s}_0) = C(\mathbf{s}_0, \mathbf{s}_0) - \mathbf{c}_Y(\mathbf{s}_0)' \mathbf{C}_Z^{-1} \mathbf{c}_Y(\mathbf{s}_0).$$

The simple kriging MSPE is a function of the design points,  $\mathbf{D} = (\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_N)$ , through  $\mathbf{c}_Y(\mathbf{s}_0)$  and  $\mathbf{C}_Z^{-1}$ .

From a Bayesian perspective, this predictor can be rationalized by a particular Bayesian hierarchical model (HM) that has a bit more structure. Specifically, we now assume that  $\{Y(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$  and  $\{\varepsilon(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$  are independent Gaussian processes. Then the posterior predictive distribution for  $Y(\mathbf{s}_0)$  can be derived as

$$[Y(\mathbf{s}_0)|\mathbf{Z}] = \int [Y(\mathbf{s}_0)|\mathbf{Y}, \mathbf{Z}][\mathbf{Y}|\mathbf{Z}]d\mathbf{Y}$$

where it is easy to derive that

$$\begin{aligned} Y(\mathbf{s}_0)|\mathbf{Y}, \mathbf{Z} &\sim N [\mathbf{x}(\mathbf{s}_0)'\boldsymbol{\beta} + \mathbf{c}_Y(\mathbf{s}_0)'\mathbf{C}_Y^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}), C(\mathbf{s}_0, \mathbf{s}_0) - \mathbf{c}_Y(\mathbf{s}_0)'\mathbf{C}_Y^{-1}\mathbf{c}_Y(\mathbf{s}_0)], \\ \mathbf{Y}|\mathbf{Z} &\sim N [\mathbf{X}\boldsymbol{\beta} + \mathbf{C}_Y\mathbf{C}_Z^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}), \mathbf{C}_Y - \mathbf{C}_Y\mathbf{C}_Z^{-1}\mathbf{C}_Y] \end{aligned}$$

so that  $Y(\mathbf{s}_0)|\mathbf{Z}$  is Gaussian with

$$\begin{aligned} E[Y(\mathbf{s}_0)|\mathbf{Z}] &= E\{E[Y(\mathbf{s}_0)|\mathbf{Y}, \mathbf{Z}]\mathbf{Z}\} = \mathbf{x}(\mathbf{s}_0)'\boldsymbol{\beta} + \mathbf{c}_Y(\mathbf{s}_0)'\mathbf{C}_Y^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}), \\ \text{var}[Y(\mathbf{s}_0)|\mathbf{Z}] &= E\{\text{var}[Y(\mathbf{s}_0)|\mathbf{Y}, \mathbf{Z}]\mathbf{Z}\} + \text{var}\{E[Y(\mathbf{s}_0)|\mathbf{Y}, \mathbf{Z}]\mathbf{Z}\} \\ &= C(\mathbf{s}_0, \mathbf{s}_0) - \mathbf{c}_Y(\mathbf{s}_0)'\mathbf{C}_Y^{-1}\mathbf{c}_Y(\mathbf{s}_0). \end{aligned}$$

Then a design criterion we could minimize is average simple kriging MSPE across all target locations,  $\sum_{i=1}^N \sigma_{sk}^2(\mathbf{s}_i)/N$ . This is a bit naive, however, because it does not take into account the covariance structure across the locations.

Let  $\mathbf{Y}^* = [Y(\mathbf{s}_1), Y(\mathbf{s}_2), \dots, Y(\mathbf{s}_M)]'$  denote the true vector of response variables at the target locations,  $\{\mathbf{s}_k \in \mathcal{D} : k = 1, \dots, M\}$ . The covariance between the prediction errors at  $\mathbf{s}_i$  and  $\mathbf{s}_j$  can be derived as  $\text{cov}[Y(\mathbf{s}_i), Y(\mathbf{s}_j)|\mathbf{Z}] = C(\mathbf{s}_i, \mathbf{s}_j) - \mathbf{c}_Y(\mathbf{s}_i)'\mathbf{C}_Z^{-1}\mathbf{c}_Y(\mathbf{s}_j)$ . Let

$\mathbf{C}_{Y^*} = \text{cov}(\mathbf{Y}^*)$  where  $\text{cov}[Y(\mathbf{s}_i), Y(\mathbf{s}_j)] = C(\mathbf{s}_i, \mathbf{s}_j)$  and  $\mathbf{C}_{YY^*} = \text{cov}(\mathbf{Y}, \mathbf{Y}^*)$  where the  $i, j$ th element is given by  $\text{cov}[Y(\mathbf{d}_i), Y(\mathbf{s}_j)] = C(\mathbf{d}_i, \mathbf{s}_j)$  for  $i = 1, 2, \dots, N$  and  $j = 1, 2, \dots, M$ . Then the simple kriging error covariance matrix is given by  $\Sigma_{sk} = \mathbf{C}_{Y^*} - \mathbf{C}'_{YY^*} \mathbf{C}_Z^{-1} \mathbf{C}_{YY^*}$ , which corresponds exactly to the posterior predictive covariance matrix for  $\mathbf{Y}^*$ .

An intuitive design criterion to minimize is  $\log |\Sigma_{sk}|$ , which is in some sense the total amount of variation in the distribution of the errors from predicting  $\mathbf{Y}^*$ . Let  $\mathbf{X}^* = [\mathbf{x}(\mathbf{s}_1), \mathbf{x}(\mathbf{s}_2), \dots, \mathbf{x}(\mathbf{s}_M)]'$ . Then if we return to the Bayesian setting  $\mathbf{Y}^*|\mathbf{Z}$  is Gaussian with mean  $\hat{\mathbf{Y}}_{sk} = [\hat{Y}_{sk}(\mathbf{s}_1), \hat{Y}_{sk}(\mathbf{s}_2), \dots, \hat{Y}_{sk}(\mathbf{s}_M)]'$  and covariance matrix  $\Sigma_{sk}$ . So if we instead maximize the expected entropy gain, we maximize

$$\begin{aligned} \mathbb{E}\{\log[\mathbf{Y}^*|\mathbf{Z}]\} &= -\frac{M}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_{sk}| - \frac{1}{2} \mathbb{E} \left\{ \mathbb{E}[(\mathbf{Y}^* - \hat{\mathbf{Y}}_{sk})' \Sigma_{sk}^{-1} (\mathbf{Y}^* - \hat{\mathbf{Y}}_{sk}) | \mathbf{Z}] \right\} \\ &= -\frac{M}{2} \log(2\pi e) - \frac{1}{2} \log |\Sigma_{sk}|. \end{aligned}$$

Thus maximizing expected entropy gain in the distribution of  $\mathbf{Y}^*$  is equivalent to minimizing  $\log |\Sigma_{sk}|$ , our intuitive design criterion.

## 1.2 Universal Kriging

A major limitation of simple kriging is that typically  $\beta$ ,  $\sigma_\epsilon^2$ , and  $C(\cdot, \cdot)$  are unknown. Universal kriging attempts to remedy this by allowing  $\beta$  to be unknown. The universal kriging predictor is

$$\hat{Y}_{uk}(\mathbf{s}_0) = \mathbf{x}(\mathbf{s}_0)' \hat{\beta}_{gls} + \mathbf{c}_Y(\mathbf{s}_0)' \mathbf{C}_Z^{-1} (\mathbf{Z} - \mathbf{X} \hat{\beta}_{gls})$$

where  $\hat{\beta}_{gls} = (\mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{Z}$  is the generalized least squares estimate of  $\beta$ , and the MSPE of  $\hat{Y}_{uk}(\mathbf{s}_0)$  is

$$\begin{aligned} \sigma_{uk}^2(\mathbf{s}_0) &= C(\mathbf{s}_0, \mathbf{s}_0) - \mathbf{c}_Y(\mathbf{s}_0)' \mathbf{C}_Z^{-1} \mathbf{c}_Y(\mathbf{s}_0) \\ &\quad + [\mathbf{x}(\mathbf{s}_0) - \mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{c}_Y(\mathbf{s}_0)]' [\mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{X}] [\mathbf{x}(\mathbf{s}_0) - \mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{c}_Y(\mathbf{s}_0)]. \end{aligned}$$

This can also be justified via a Bayesian HM by using the same Gaussian assumptions as before, but further assuming that  $\boldsymbol{\beta}$  has the improper uniform prior, i.e.  $\boldsymbol{\beta} \sim U(-\infty, \infty)$ . We will take a slightly more general approach as assume that  $\boldsymbol{\beta} \sim N(\mathbf{b}, \mathbf{S})$  where the improper uniform prior results from  $\mathbf{S} = s^2 \mathbf{I}$  as  $s \rightarrow \infty$ . Let  $\mathbf{C}_Y = \text{cov}(\mathbf{Y})$  where  $\text{cov}(Y(\mathbf{d}_i), Y(\mathbf{d}_j)) = C(\mathbf{d}_i, \mathbf{d}_j)$ . Now we have

$$[Y(\mathbf{s}_0)|\mathbf{Z}] = \int \int [Y(\mathbf{s}_0)|\mathbf{Y}, \boldsymbol{\beta}, \mathbf{Z}][\mathbf{Y}|\mathbf{Z}] d\mathbf{Y} d\boldsymbol{\beta}$$

where

$$\begin{aligned} Y(\mathbf{s}_0)|\mathbf{Y}, \boldsymbol{\beta}, \mathbf{Z} &\sim N[\mathbf{x}(\mathbf{s}_0)' \boldsymbol{\beta} + \mathbf{c}_Y(\mathbf{s}_0)' \mathbf{C}_Y^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}), C(\mathbf{s}_0, \mathbf{s}_0) - \mathbf{c}_Y(\mathbf{s}_0)' \mathbf{C}_Y^{-1} \mathbf{c}_Y(\mathbf{s}_0)], \\ \mathbf{Y}|\boldsymbol{\beta}, \mathbf{Z} &\sim N[\mathbf{X}\boldsymbol{\beta} + \mathbf{C}_Y \mathbf{C}_Z^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}), \mathbf{C}_Y - \mathbf{C}_Y \mathbf{C}_Z^{-1} \mathbf{C}_Y], \\ \boldsymbol{\beta}|\mathbf{Z} &\sim N[\hat{\boldsymbol{\beta}}_{bgl}, (\mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{X} + \mathbf{S}^{-1})^{-1}], \end{aligned}$$

where  $\hat{\boldsymbol{\beta}}_{bgl} = (\mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{X} + \mathbf{S}^{-1})^{-1}(\mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{Z} + \mathbf{S}^{-1} \mathbf{b})$  can be thought of as the Bayesian generalized least squares estimator. Then  $Y(\mathbf{s}_0)|\mathbf{Z}$  is again Gaussian with

$$\begin{aligned} E[Y(\mathbf{s}_0)|\mathbf{Z}] &= E\{E[Y(\mathbf{s}_0)|\mathbf{Y}, \boldsymbol{\beta}, \mathbf{Z}]\mathbf{Z}\} = \mathbf{x}(\mathbf{s}_0)' \hat{\boldsymbol{\beta}}_{bgl} + \mathbf{c}_Y(\mathbf{s}_0)' \mathbf{C}_Z^{-1}(\mathbf{Z} - \mathbf{X} \hat{\boldsymbol{\beta}}_{bgl}), \\ \text{var}[Y(\mathbf{s}_0)|\mathbf{Z}] &= E\{\text{var}[Y(\mathbf{s}_0)|\mathbf{Y}, \boldsymbol{\beta}, \mathbf{Z}]\mathbf{Z}\} + \text{var}\{E[Y(\mathbf{s}_0)|\mathbf{Y}, \boldsymbol{\beta}, \mathbf{Z}]\mathbf{Z}\} \\ &= C(\mathbf{s}_0, \mathbf{s}_0) - \mathbf{c}_Y(\mathbf{s}_0)' \mathbf{C}_Y^{-1} \mathbf{c}_Y(\mathbf{s}_0) + \text{var}\{[\mathbf{x}(\mathbf{s}_0) - \mathbf{X}' \mathbf{C}_Y^{-1} \mathbf{c}_Y(\mathbf{s}_0)]' \boldsymbol{\beta} + \mathbf{c}_Y(\mathbf{s}_0)' \mathbf{C}_Y^{-1} \mathbf{Y}|\mathbf{Z}\}. \end{aligned}$$

To compute the last variance term, note that

$$\begin{aligned} \text{cov}(\mathbf{Y}|\mathbf{Z}) &= E[\text{cov}(\mathbf{Y}|\boldsymbol{\beta}, \mathbf{Z})|\mathbf{Z}] + \text{cov}[E(\mathbf{Y}|\boldsymbol{\beta}, \mathbf{Z})|\mathbf{Z}] \\ &= \mathbf{C}_Y - \mathbf{C}_Y \mathbf{C}_Z^{-1} \mathbf{C}_Y + (\mathbf{I} - \mathbf{C}_Y \mathbf{C}_Z^{-1}) \mathbf{X} (\mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{X} + \mathbf{S}^{-1})^{-1} \mathbf{X}' (\mathbf{I} - \mathbf{C}_Z^{-1} \mathbf{C}_Y) \end{aligned}$$

and

$$\begin{aligned} \text{cov}(\mathbf{Y}, \boldsymbol{\beta}|\mathbf{X}) &= E[\text{cov}(\mathbf{Y}, \boldsymbol{\beta}|\boldsymbol{\beta}, \mathbf{Z})|\mathbf{Z}] + \text{cov}[E(\mathbf{Y}|\boldsymbol{\beta}, \mathbf{Z}), \boldsymbol{\beta}|\mathbf{Z}] \\ &= (\mathbf{I} - \mathbf{C}_Y \mathbf{C}_Z^{-1}) \mathbf{X} (\mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{X} + \mathbf{S}^{-1})^{-1}. \end{aligned}$$

Then we have

$$\begin{aligned}
\text{var}[Y(\mathbf{s}_0)|\mathbf{Z}] = & \\
& C(\mathbf{s}_0, \mathbf{s}_0) - \mathbf{c}_Y(\mathbf{s}_0)' \mathbf{C}_Y^{-1} \mathbf{c}_Y(\mathbf{s}_0) \\
& + [\mathbf{x}(\mathbf{s}_0) - \mathbf{X}' \mathbf{C}_Y^{-1} \mathbf{c}_Y(\mathbf{s}_0)]' [\mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{X} + \mathbf{S}^{-1}]^{-1} [\mathbf{x}(\mathbf{s}_0) - \mathbf{X}' \mathbf{C}_Y^{-1} \mathbf{c}_Y(\mathbf{s}_0)] \\
& + \mathbf{c}_Y(\mathbf{s}_0)' \{ \mathbf{C}_Y^{-1} - \mathbf{C}_Z^{-1} + (\mathbf{C}_Y^{-1} - \mathbf{C}_Z^{-1}) \mathbf{X} (\mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{X} + \mathbf{S}^{-1})^{-1} \mathbf{X}' (\mathbf{C}_Y^{-1} - \mathbf{C}_Z^{-1}) \} \mathbf{c}_Y(\mathbf{s}_0) \\
& + \mathbf{c}_Y(\mathbf{s}_0)' (\mathbf{C}_Y^{-1} - \mathbf{C}_Z^{-1}) \mathbf{X} (\mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{X} + \mathbf{S}^{-1})^{-1} [\mathbf{x}(\mathbf{s}_0) - \mathbf{X}' \mathbf{C}_Y^{-1} \mathbf{c}_Y(\mathbf{s}_0)] \\
& + [\mathbf{x}(\mathbf{s}_0) - \mathbf{X}' \mathbf{C}_Y^{-1} \mathbf{c}_Y(\mathbf{s}_0)]' (\mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{X} + \mathbf{S}^{-1})^{-1} \mathbf{x}(\mathbf{s}_0)' (\mathbf{C}_Y^{-1} - \mathbf{C}_Z^{-1}) \mathbf{c}_Y(\mathbf{s}_0),
\end{aligned}$$

which reduces to

$$\begin{aligned}
\text{var}[Y(\mathbf{s}_0)|\mathbf{Z}] = & C(\mathbf{s}_0, \mathbf{s}_0) - \mathbf{c}_Y(\mathbf{s}_0)' \mathbf{C}_Z^{-1} \mathbf{c}_Y(\mathbf{s}_0) \\
& + [\mathbf{x}(\mathbf{s}_0) - \mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{c}_Y(\mathbf{s}_0)]' [\mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{X} + \mathbf{S}^{-1}]^{-1} [\mathbf{x}(\mathbf{s}_0) - \mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{c}_Y(\mathbf{s}_0)].
\end{aligned}$$

If  $\mathbf{S} = s\mathbf{I}$ , then as  $s \rightarrow \infty$ ,  $\mathbf{S}^{-1} \rightarrow 0 \times \mathbf{I}$ , so  $\mathbf{S}^{-1}$  drops out of all of these equations and  $\hat{\boldsymbol{\beta}}_{bgl s} \rightarrow \hat{\boldsymbol{\beta}}_{gls}$ , so that the posterior predictive mean is the same as the universal kriging predictor and the posterior predictive variance is the same as the variance of the universal kriging prediction error.

The covariance between the kriging prediction errors at two locations is

$$\begin{aligned}
\text{cov}[\hat{Y}_{uk}(\mathbf{s}_i), \hat{Y}_{uk}(\mathbf{s}_j)] = & C(\mathbf{s}_i, \mathbf{s}_j) - \mathbf{c}_Y(\mathbf{s}_i)' \mathbf{C}_Z^{-1} \mathbf{c}_Y(\mathbf{s}_j) \\
& + [\mathbf{x}(\mathbf{s}_i) - \mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{c}_Y(\mathbf{s}_i)]' [\mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{X}]^{-1} [\mathbf{x}(\mathbf{s}_j) - \mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{c}_Y(\mathbf{s}_j)]
\end{aligned}$$

and similarly, the posterior predictive covariance between two locations can be derived as

$$\begin{aligned}
\text{cov}[Y(\mathbf{s}_i), Y(\mathbf{s}_j)|\mathbf{Z}] = & C(\mathbf{s}_i, \mathbf{s}_j) - \mathbf{c}_Y(\mathbf{s}_i)' \mathbf{C}_Z^{-1} \mathbf{c}_Y(\mathbf{s}_j) \\
& + [\mathbf{x}(\mathbf{s}_i) - \mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{c}_Y(\mathbf{s}_i)]' [\mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{X} + \mathbf{S}^{-1}]^{-1} [\mathbf{x}(\mathbf{s}_j) - \mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{c}_Y(\mathbf{s}_j)]
\end{aligned}$$

and once again the latter goes to the former as  $s \rightarrow \infty$  when  $\mathbf{S} = s\mathbf{I}$ . So  $\mathbf{Y}^*$  is Gaussian in

the posterior predictive distribution with mean and covariance matrix

$$\begin{aligned}\hat{\mathbf{Y}}_{buk} &= \mathbb{E}[\mathbf{Y}^*|\mathbf{Z}] = \mathbf{X}^* \hat{\boldsymbol{\beta}}_{bgl s} + \mathbf{C}'_{YY^*} \mathbf{C}_Z^{-1} (\mathbf{Z} - \mathbf{X} \hat{\boldsymbol{\beta}}_{bgl s}) \\ \boldsymbol{\Sigma}_{buk} &= \mathbf{C}_{Y^*} - \mathbf{C}'_{YY^*} \mathbf{C}_Z^{-1} \mathbf{C}_{YY^*} \\ &\quad + (\mathbf{X}^* - \mathbf{C}'_{YY^*} \mathbf{C}_Z^{-1} \mathbf{X}) (\mathbf{X}' \mathbf{C}_Z^{-1} \mathbf{X} + \mathbf{S}^{-1})^{-1} (\mathbf{X}^* - \mathbf{C}'_{YY^*} \mathbf{C}_Z^{-1} \mathbf{X})'.\end{aligned}$$

Minimizing  $\log |\boldsymbol{\Sigma}_{buk}|$  in order to choose the sampled locations in this model seems intuitive, much like in the simple kriging case. Again this corresponds to maximizing the expected entropy gain. The part of expected entropy gain that depends on the design is

$$\begin{aligned}\mathbb{E} \{ \log[\mathbf{Y}^*|\mathbf{Z}] \} &= -\frac{M}{2} - \frac{1}{2} \log |\boldsymbol{\Sigma}_{uk}| - \frac{1}{2} \mathbb{E} \left\{ \mathbb{E}[(\mathbf{Y}^* - \hat{\mathbf{Y}}_{uk})' \boldsymbol{\Sigma}_{sk}^{-1} (\mathbf{Y}^* - \hat{\mathbf{Y}}_{uk}) | \mathbf{Z}] \right\} \\ &= -\frac{M}{2} \log(2\pi e) - \frac{1}{2} \log |\boldsymbol{\Sigma}_{buk}|,\end{aligned}$$

and maximizing this is equivalent to minimizing  $\log |\boldsymbol{\Sigma}_{buk}|$ . Since  $\boldsymbol{\Sigma}_{buk} = \boldsymbol{\Sigma}_{uk}$  when  $\boldsymbol{\beta} \sim U(-\infty, \infty)$ , this includes the case of minimizing the log determinant of the covariance of the universal kriging prediction errors.

### 1.3 Empirical Kriging

In both simple and universal kriging the predictors and their MSPEs are the same as the posterior predictive means and posterior predictive variances for a suitably chosen Bayesian HM, and in both cases if we wish to choose the sampled locations in order to minimize the log determinant of the covariance matrix of the kriging prediction errors, the problem is equivalent to maximizing the expected entropy gain for the predictive distribution at the target locations. We can easily the entropy method to the case where both  $\sigma_\varepsilon^2$  and  $C(.,.)$  are unknown, called empirical kriging, though it may no longer correspond to the optimal linear predictor. In the non-Bayesian case things are a bit more delicate.

Suppose the covariance function is parameterized by  $\boldsymbol{\theta} \in \Theta$ , denoted by  $C_{\boldsymbol{\theta}}(.,.)$ , and let  $\boldsymbol{\phi} = (\boldsymbol{\theta}, \sigma_\varepsilon^2)$ . Let  $\mathcal{I}(\boldsymbol{\phi})$  denote the Fisher information matrix for estimating  $\boldsymbol{\phi}$ , e.g.

from maximum likelihood and  $\hat{\phi}$  the corresponding estimate. Then the empirical kriging estimate of  $Y(\mathbf{s}_0)$  is given by  $\hat{Y}_{ek}(\mathbf{s}_0) = \hat{Y}_{uk}(\mathbf{s}_0; \hat{\phi})$ , where we now make the dependence of  $\hat{Y}_{uk}$  on  $\phi$  explicit. The MSPE of  $\hat{Y}_{ek}$  can be approximated by  $\sigma_{ek}^2(\mathbf{s}_0; \phi) \approx \sigma_{uk}^2(\mathbf{s}_0; \phi) + \text{tr}[A(\mathbf{s}_0; \phi)\mathcal{I}^{-1}(\phi)]$  where  $A(\mathbf{s}_0; \phi) = \text{var}[\partial\hat{Y}_{uk}(\mathbf{s}_0; \phi)/\partial\phi]$ . [THIS WILL BE HARD TO DERIVE.... DO WE EVEN WANT TO?]

In the Bayesian case, assume that a priori  $(\beta, \phi) \sim [\beta, \phi]$  — for now we allow  $\beta$  to have an arbitrary prior distribution. The part of the expected entropy gain for the predictive distribution that depends on the design is

$$\begin{aligned} \mathbb{E} \{\log[\mathbf{Y}^*|\mathbf{Z}]\} &= \int \int \log[\mathbf{Y}^*|\mathbf{Z}][\mathbf{Y}^*, \mathbf{Z}]d\mathbf{Y}^*d\mathbf{Z} \\ &= \int \int \log[\mathbf{Y}^*|\mathbf{Z}] \int [\mathbf{Y}^*, \mathbf{Z}|\beta, \phi][\beta, \phi]d\mathbf{Y}^*d\mathbf{Z}d\beta d\phi. \end{aligned}$$

This quantity is not available in closed form, so we estimate it via a hybrid of Monte Carlo simulation and marginal likelihood estimation.

Typically,  $[\beta, \phi]$  is easy to simulate from, and  $[\mathbf{Y}^*, \mathbf{Z}|\beta, \phi]$  is Gaussian, so it is easy to obtain  $L$  draws from  $[\mathbf{Y}^*, \mathbf{Z}]$ , denoted by  $[(\mathbf{Y}^*)^{(l)}, \mathbf{Z}^{(l)}]$  for  $l = 1, 2, \dots, L$ . Next, we need an estimator of  $\log[\mathbf{Y}^*|\mathbf{Z}]$ . We can rewrite the density as  $[\mathbf{Y}^*|\mathbf{Z}] = [\mathbf{Y}^*, \mathbf{Z}]/[\mathbf{Z}]$ . Both the numerator and demoninator are marginal likelihoods, and there are numerous ways to estimate these in the literature. In the typical Bayesian design case, we are trying to maximize the expected entropy gain of the model parameter, i.e. of  $[\beta, \phi|\mathbf{Z}]$ . In that case there is the joint distribution  $[\mathbf{Z}, \beta, \phi]$ , and is available in closed form. Since we are interested in the entropy of the predictive distribution at a fixed set of points, this becomes more complicated.

When  $[\beta, \phi] = [\beta][\phi]$  with  $\beta \sim N(\mathbf{b}, \mathbf{S})$ , the marginal likelihood estimation problems become easier. Then we have  $[\mathbf{Y}^*|\mathbf{Z}, \phi]$  and  $[\mathbf{Z}|\phi]$  in closed form from the universal kriging



section, and

$$[\mathbf{Y}^*, \mathbf{Z}] = \int [\mathbf{Y}^* | \mathbf{Z}, \phi] [\mathbf{Z} | \phi] [\phi] d\phi \quad \text{and} \quad [\mathbf{Z}] = \int [\mathbf{Z} | \phi] [\phi] d\phi.$$

These integrals are still marginal likelihoods, but the dimension of  $\phi$  is often small enough that numerical integration may be feasible.