Bi-fidelity Surrogate Modelling: Showcasing the need for new test instances - Appendix

Nicolau Andrés-Thió, Mario Andrés Muñoz, Kate Smith-Miles

Appendix A: Kriging

Kriging was developed by Danie Gerhardus Krige empirically to evaluate mineral resources (Krige, 1951); his work was later formalised by Matheron (1963) and has been expanded on by many scientists. Jones (2001) presents a gentle introduction to Kriging and some of its basic uses; the standard derivation can be seen in Sacks et al. (1989) among others. Kriging is a method developed for (single-source) EBB problems, however it can be applied to Bf-EBB problems by working only with the function f_h. Therefore, in this formulation, f_h is simply denoted by f.

The formulation given by Kriging assumes the function samples made so far at locations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are realisations of random normal variables $Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_n)$ with mean μ and variance σ^2. Further, the errors are correlated based on the distance between variables, that is

$$Corr(Y(\mathbf{x}_i), Y(\mathbf{x}_j)) = exp\left\{-\sum_{k=1}^{d} \theta_k \|\mathbf{x}_i^k - \mathbf{x}_j^k\|^{p_k}\right\}$$

thus, the multivariate random variable $\mathbf{Y} = [Y(\mathbf{x}_1) \ldots Y(\mathbf{x}_n)]$ has the distribution $\mathbf{Y} \sim N(\mathbf{1}\mu, \sigma^2 R)$, with $R_{i,j} = exp\left\{-\sum_{k=1}^{d} \theta_k \|\mathbf{x}_i^k - \mathbf{x}_j^k\|^{p_k}\right\}$. Note this has the hyperparameters $\mu, \sigma^2, \theta_1, \ldots, \theta_d, p_1, \ldots, p_d$. The values θ_k and p_k give an indication of the effect of moving along any of the dimensions (i.e changing the value of a single variable). The constant θ_k represents how the correlation changes with distance: small values mean there is no correlation even for close points in the k^{th} dimension, but large values indicate even relatively distant sample points (in the k^{th} dimension) are correlated. The constant p_k allows the technique to model from smooth functions ($p_k = 2$) to rough, non-differentiable ones ($p_k \to 0$).

In order to fit the model to the data, the log density function of $\mathbf{Y}$ is maximised, which after simplification leads to the auxiliary optimisation problem

$$\max_{\theta_1,\ldots,\theta_d,p_1,\ldots,p_d} -\frac{n}{2}\log(\hat{\sigma}^2) - \frac{1}{2}\log(|R|)$$

with

$$\begin{aligned} \hat{\mu} &= \frac{\mathbf{1}^T R^{-1}\mathbf{y}}{\mathbf{1}^T R^{-1}\mathbf{1}} \\ \hat{\sigma}^2 &= \frac{(\mathbf{y} - \mathbf{1}\hat{\mu})^T R^{-1}(\mathbf{y} - \mathbf{1}\hat{\mu})}{n} \\ \mathbf{y} &= [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)] \end{aligned}$$

Note this problem cannot be solved analytically and thus the tuning of these hyperparameters is an auxiliary problem that must be solved. Once the model has been trained, for a given sample point $\mathbf{x}$ Kriging provides the most likely objective function $s(\mathbf{x})$ and the variance of the estimate $v^2(\mathbf{x})$. These are given by

$$\begin{aligned} s(\mathbf{x}) &= \hat{\mu} + \mathbf{r}^T R^{-1}(\mathbf{y} - \mathbf{1}\hat{\mu}) \\ v^2(\mathbf{x}) &= \hat{\sigma}^2 \left[1 - \mathbf{r}^T R^{-1}\mathbf{r} + \frac{(1 - \mathbf{1}^T R^{-1}\mathbf{r})^2}{\mathbf{1}^T R^{-1}\mathbf{1}}\right] \end{aligned}$$

where

$$\mathbf{r} = \begin{bmatrix} Corr(Y(\mathbf{x}), Y(\mathbf{x}_1)) \\ \dots \\ Corr(Y(\mathbf{x}), Y(\mathbf{x}_n)) \end{bmatrix}$$

Thus, Kriging thinks of the objective function predictor as being a realisation of $Y(\mathbf{x}) \sim N(s(\mathbf{x}), v^2(\mathbf{x}))$. This allows the asking of questions such as what is the uncertainty or what is the expected improvement at a sample point, among others.

Appendix B: Co-Kriging

Kennedy and O'Hagan (2000) present a technique to use multiple information sources combined with Gaussian processes for global optimisation. This technique can be adapted to Kriging (Forrester et al., 2007), producing a new technique known as Co-Kriging. Similarly to Kriging, the idea behind Co-Kriging is to model the responses of the cheap objective function f_l at sample points $\mathbf{X}_l = (\mathbf{x}_1^l, \mathbf{x}_2^l, \dots, \mathbf{x}_{n_l}^l)$ and the responses of the expensive objective function f_h at sample points $\mathbf{X}_h = (\mathbf{x}_1^h, \mathbf{x}_2^h, \dots, \mathbf{x}_{n_h}^h)$ as the realisation of a multivariate random variable:

$$\mathbf{Y} = (\mathbf{Y}_l(\mathbf{X}_l), \mathbf{Y}_h(\mathbf{X}_h)) = (Y_l(\mathbf{x}_1^l), \dots, Y_l(\mathbf{x}_{n_l}^l), Y_h(\mathbf{x}_1^h), \dots, Y_h(\mathbf{x}_{n_h}^h))$$

The multivariate random variable $\mathbf{Y}_l(\mathbf{X}_l)$, that is the response of the cheap objective function, is treated as a multivariate normal random variable with distribution $N(\mu_l, \sigma_l^2 R_l)$. On the other hand, the multivariate random variable $\mathbf{Y}_h(\mathbf{X}_h)$, that is the response of the expensive objective function, is represented by a scaling of ρ of the response of the cheap expensive function $\mathbf{Y}_l(\mathbf{X}_l)$ plus a new Gaussian process $\mathbf{Y}_b$ which models the difference between the cheap and expensive objective functions, that is

$$\mathbf{Y}_h(\mathbf{X}_h) = \rho \mathbf{Y}_l(\mathbf{X}_h) + \mathbf{Y}_b(\mathbf{X}_h)$$

The random variable $\mathbf{Y}_b$ is also treated as a multivariate normal random variable with distribution $N(\mu_b, \sigma_b^2 R_b)$. It is assumed that $\mathbf{Y}_l$ and $\mathbf{Y}_b$ are independent. Thus the following correlation measures are given for $\mathbf{Y}_l$ and $\mathbf{Y}_h$:

$$
\begin{aligned}
Corr(\mathbf{Y}_l(\mathbf{X}_l), \mathbf{Y}_l(\mathbf{X}_l)) &= \sigma_l^2 R_l(\mathbf{X}_l, \mathbf{X}_l) \\
Corr(\mathbf{Y}_h(\mathbf{X}_h), \mathbf{Y}_l(\mathbf{X}_l)) &= \rho\sigma_l^2 R_l(\mathbf{X}_h, \mathbf{X}_l) \\
Corr(\mathbf{Y}_h(\mathbf{X}_h), \mathbf{Y}_h(\mathbf{X}_h)) &= \rho^2\sigma_l^2 R_l(\mathbf{X}_h, \mathbf{X}_h) + \sigma_b^2 R_b(\mathbf{X}_h, \mathbf{X}_h)
\end{aligned}
$$

where

$$
\begin{aligned}
R_l(\mathbf{X}_l, \mathbf{X}_l)_{i,j} &= exp\left\{ -\sum_{k=1}^{d} \theta_k^l \|(\mathbf{x}_i^l)_k - (\mathbf{x}_j^l)_k\|^{p_k^l} \right\} \qquad 1 \le i,j \le n_l \\
R_l(\mathbf{X}_h, \mathbf{X}_l)_{i,j} &= exp\left\{ -\sum_{k=1}^{d} \theta_k^l \|(\mathbf{x}_i^h)_k - (\mathbf{x}_j^l)_k\|^{p_k^l} \right\} \qquad 1 \le i \le n_h \quad 1 \le j \le n_l \\
R_l(\mathbf{X}_h, \mathbf{X}_h)_{i,j} &= exp\left\{ -\sum_{k=1}^{d} \theta_k^l \|(\mathbf{x}_i^h)_k - (\mathbf{x}_j^h)_k\|^{p_k^l} \right\} \qquad 1 \le i,j \le n_h \\
R_b(\mathbf{X}_h, \mathbf{X}_h)_{i,j} &= exp\left\{ -\sum_{k=1}^{d} \theta_k^b \|(\mathbf{x}_i^h)_k - (\mathbf{x}_j^h)_k\|^{p_k^b} \right\} \qquad 1 \le i,j \le n_h
\end{aligned}
$$

In order to fit the model to the data, the log density function of $\mathbf{Y}_l$ is maximised, which after simplification (Forrester et al., 2007) leads to the auxiliary optimisation problem

$$
\max_{\mu_l, \sigma_l^2, \theta_1^l, \ldots, \theta_d^l, p_1^l, \ldots, p_d^l} -\frac{n_l}{2}\log(\hat{\sigma}_l^2) - \frac{1}{2}\log(|det(R_l(\mathbf{X}_l, \mathbf{X}_l))|)
$$

where

$$
\begin{aligned}
\hat{\mu}_l &= \frac{\mathbf{1}^T R_l(\mathbf{X}_l, \mathbf{X}_l)^{-1}\mathbf{Y}_l}{\mathbf{1}^T R_l(\mathbf{X}_l, \mathbf{X}_l)^{-1}\mathbf{1}} \\
\hat{\sigma}_l^2 &= \frac{(\mathbf{Y}_l - \mathbf{1}\hat{\mu}_l)^T R_l(\mathbf{X}_l, \mathbf{X}_l)^{-1}(\mathbf{Y}_l - \mathbf{1}\hat{\mu}_l)}{n_l} \\
\mathbf{Y}_l &= (f_l(\mathbf{x}_1^l), \ldots, f_l(\mathbf{x}_{n_l}^l))
\end{aligned}
$$

As is the case with Kriging, this problem cannot be solved analytically and is an auxiliary optimisation problem which must be solved. In order to calculate the parameters associated with $\mathbf{Y}_b$, first $\mathbf{b}$ is defined:

$$
\mathbf{b} = \mathbf{Y}_h - \rho\mathbf{Y}_l(\mathbf{X}_h)
$$

where $\mathbf{Y}_h = (f_h(\mathbf{x}_1^h), \ldots, f_h(\mathbf{x}_{n_h}^h))$, and $\mathbf{Y}_l(\mathbf{X}_h)_i$ is $f_l(\mathbf{x}_i^h)$ if the point has already been evaluated, and otherwise it is $\hat{y}(\mathbf{x}_i^h) = \hat{\mu}_l + \mathbf{r}_l^T R_l(\mathbf{X}_l, \mathbf{X}_l)^{-1}(\mathbf{Y}_l - \mathbf{1}\hat{\mu}_l)$, with $\mathbf{r}_l = (R_l(\mathbf{x}, \mathbf{x}_1^l), \ldots, R_l(\mathbf{x}, \mathbf{x}_{n_l}^l))$. That is, if a point has not been evaluated by f_l yet, its Kriging predictor of the cheap model is used instead. A second auxiliary problem is solved to find a second set of hyperparameters, using the log density function of $\mathbf{Y}_b$:

$$\max_{\rho,\theta_1^b,\ldots,\theta_d^b,p_1^b,\ldots,p_d^b} -\frac{n_h}{2}\log(\hat{\sigma}_b^2) - \frac{1}{2}\log(|det(R_b(\mathbf{X}_h,\mathbf{X}_h))|)$$

where

$$\begin{aligned}
\hat{\mu}_b &= \frac{\mathbf{1}^T R_b(\mathbf{X}_h,\mathbf{X}_h)^{-1}\mathbf{b}}{\mathbf{1}^T R_b(\mathbf{X}_h,\mathbf{X}_h)^{-1}\mathbf{1}} \\
\hat{\sigma}_b^2 &= \frac{(\mathbf{b}-\mathbf{1}\hat{\mu}_b)^T R_b(\mathbf{X}_h,\mathbf{X}_h)^{-1}(\mathbf{b}-\mathbf{1}\hat{\mu}_b)}{n_h}
\end{aligned}$$

Finally, the Co-Kriging predictor is given by

$$s_h(\mathbf{x}) = \hat{\mu} + \mathbf{c}^T C^{-1}(\mathbf{y}-\mathbf{1}\hat{\mu})$$

$$\text{with} \qquad C = \begin{bmatrix} \hat{\sigma}_l^2 R_l(\mathbf{X}_l,\mathbf{X}_l) & \hat{\rho}\hat{\sigma}_l^2 R_l(\mathbf{X}_l,\mathbf{X}_h) \\ \hat{\rho}\hat{\sigma}_l^2 R_l(\mathbf{X}_h,\mathbf{X}_l) & \hat{\rho}^2\hat{\sigma}_l^2 R_l(\mathbf{X}_h,\mathbf{X}_h) + \hat{\sigma}_b^2 R_b(\mathbf{X}_h,\mathbf{X}_h) \end{bmatrix}$$

$$\mathbf{c} = \begin{bmatrix} \hat{\rho}\hat{\sigma}_l^2 R_l(\mathbf{X}_l,\mathbf{x}) \\ \hat{\rho}^2\hat{\sigma}_l^2 R_l(\mathbf{X}_h,\mathbf{x}) + \hat{\sigma}_b^2 R_b(\mathbf{X}_h,\mathbf{x}) \end{bmatrix}$$

$$\hat{\mu} = \frac{\mathbf{1}^T C^{-1}\mathbf{y}}{\mathbf{1}^T C^{-1}\mathbf{1}}$$

An estimated mean-square error can also be extracted for the predictor, which is given by

$$v^2(\mathbf{x}) = \hat{\rho}^2\hat{\sigma}_l^2 + \hat{\sigma}_b^2 - \mathbf{c}^T C^{-1}\mathbf{c}$$

The overall algorithm presented by Forrester consists of creating a large set of sample points for the cheap objective function, and then choose a subset of those points to sample the expensive objective function. It then chooses the next sample point by treating the value at a particular point as the realisation of a normal random variable $\sim N(s_h(\mathbf{x}), v^2(\mathbf{x}))$. The chosen sample point is evaluated both by the cheap and expensive objective functions, the models are fitted again, and the next sample point is chosen.

References

Forrester, A. I., Sóbester, A., and Keane, A. J. (2007). Multi-fidelity optimization via surrogate modelling. *Proceedings of the royal society a: mathematical, physical and engineering sciences*, 463(2088):3251–3269.

Jones, D. R. (2001). A taxonomy of global optimization methods based on response surfaces. *Journal of global optimization*, 21(4):345–383.

Kennedy, M. C. and O'Hagan, A. (2000). Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139.

Matheron, G. (1963). Principles of geostatistics. *Economic geology*, 58(8):1246–1266.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical science*, pages 409–423.