Importing Uncertainty Estimates from One Surrogate to Another

Felipe A. C. Viana* and Raphael T. Haftka†
University of Florida, Gainesville, FL, USA, 32611

In adaptive sampling and optimization methods, the uncertainty estimators are used to guide the selection of the next sampling point(s). These algorithms often limit themselves to surrogates such as kriging and polynomial response surface because of the lack of uncertainty estimates in the implementation of other surrogates. We propose the importation (borrowing) of uncertainty estimates from one surrogate to another. This would allow the use of support vector regression models together with kriging uncertainty estimates, for example. When multiple surrogates are available, we also propose using cross-validation to aid in the decision of which surrogate to import from. The approach was tested on two algebraic examples for ten basic surrogates including different instances of kriging, polynomial response surface, radial basis neural networks and support vector regression surrogates. For these examples we found that (i) the statistically based uncertainty estimates do not always correlate well to the errors; (ii) importation of uncertainty structure can offer a reasonable solution; and (iii) cross-validation successfully selects the surrogate according to the quality of the uncertainty structure and therefore it is useful in choosing which surrogate to import the uncertainty estimate from.

Nomenclature

e(x) = absolute value of the error in prediction, \( e(x) = |\hat{y}(x) - y(x)| \)
points

\( e_{RMS} \) = root mean square error

\( x \) = vector of input variables (point in the design space)

\( y(x) \) = actual response at \( x \)

\( \hat{y}(x) \) = surrogate model of \( y(x) \)

\( PRESS_{RMS} \) = square root of the mean prediction sum of errors

\( R \) = correlation between the error and the square root of the prediction variance at a large number of test points

\( R_{XV} \) = correlation estimate based on cross-validation data

\( \rho \) = ratio defined by \( \rho = \hat{\varphi}_{RMS}/e_{RMS} \)

\( \rho_{XV} \) = \( \rho \) estimate based on cross-validation data

\( \varphi_{KRG}^2(x) \) = prediction variance of a kriging model at \( x \)

\( \varphi_{PRS}^2(x) \) = prediction variance of a polynomial response surface at \( x \)

\( \varphi_{RMS} \) = estimated root mean square error based on the integration of the prediction variance

\( \varphi_{RMS_{XV}} \) = \( \varphi_{RMS} \) estimate based on cross-validation data

* Research Assistant, Dept. of Mechanical and Aerospace Engineering, fchegury@ufl.edu, Student Member AIAA.
† Distinguished Professor, Dept. of Mechanical and Aerospace Engineering, haftka@ufl.edu, Fellow AIAA.

Copyright © 2009 by Felipe A. C. Viana. Published by the American Institute of Aeronautics and Astronautics, Inc., with permission.
I. Introduction

In design optimization, surrogate models (also known as metamodels) are used to replace costly simulations and experimentations of engineering systems. Sophisticated surrogates such as kriging, radial basis neural networks, and support vector regression, increasingly share place with the traditional polynomial response surface. Besides offering cheap-to-compute prediction, in some cases (such as in kriging and in polynomial response surfaces), they also furnish estimates of the uncertainty associated with the prediction at a given point. Uncertainty estimates are used in adaptive sampling and optimization methods to select the next sampling point(s), i.e., high cost simulations. For example, the Efficient Global Optimization (EGO) and the Enhanced Sequential Optimization algorithms use the kriging prediction variance to seek the point of maximum expected improvement as the next simulation for the optimization process. While uncertainty estimates are typically available in implementations of kriging and polynomial response surfaces, but they are not commonly available for other surrogates (such as radial basis neural networks and support vector regression). Therefore, algorithms such as EGO limit themselves to kriging surrogates (and sometimes polynomial response surface).

The use of uncertainty estimates has motivated research on providing them for other surrogates. De Veaux et al., for example, applied to neural networks methods commonly used for computing prediction intervals in nonlinear regression and showed that this strategy can be effective when the number of training points is large. Seok et al. presented a Bayesian approach to estimating prediction intervals for support vector regression. Research has also been conducted in the sense of improving the estimators for kriging. For instance, Yamamoto presents an alternative error estimate for kriging that is based on a weighted average of the squared differences between data values and the retained estimate.

On the other hand, if we look at the surrogates as predictors, research has pointed to the simultaneous use of multiple surrogates rather than a single one. This makes sense because no single surrogate works well for all problems and the cost of constructing multiple surrogates is often small compared to the cost of simulations. Multiple surrogate approaches can involve: (i) fitting multiple surrogates and picking one based on cross-validation errors, and (ii) performing optimization multiple times by using multiple surrogates.

In this work, we explore the scenario in which multiple surrogates can be generated. Some of them deliver both the predictor and the error estimate; but others just the predictor. We propose that surrogates can use error estimates from other models (we called this process as importation of the error estimate). This would culminate in a situation where the predictor is given by a surrogate and the error estimate by another. We explore the following questions:

1. Is it possible to import (borrow) uncertainty estimates from one surrogate for use in another? For example, can we use the uncertainty structure of a kriging model together with a support vector regression surrogate?
2. If multiple surrogates are available, how to choose the best uncertainty structure?

The remaining of the paper is organized as follows. Section II presents the background on surrogate and uncertainty estimates. Section IV details our proposed strategy for importation of uncertainty estimates. Section V presents the results of numerical experiments and some discussion. Finally, Section VI presents concluding remarks.

II. Background

We denote by \( y \) the response of a numerical simulator or function that is to be studied:

\[
y : D \subset \mathbb{R}^d \rightarrow \mathbb{R}, \quad x \mapsto y(x),
\]

where \( x = \{x_1, \ldots, x_d\}^T \) is a \( d \)-dimensional vector of input variables.

When the response \( y(x) \) is expensive to evaluate, we may approximate it by a cheaper model \( \hat{y}(x) \) (surrogate model), based on (i) assumptions on the nature of \( y(x) \); and (ii) on the observed values of \( y(x) \) at a set of points, called the design of experiment (DOE). More explicitly:

\[
y(x) = \hat{y}(x) + \varepsilon(x),
\]

where \( \varepsilon(x) \) represents both the error of approximation and measurement (random) errors.
The absolute value of the error in the surrogate prediction is:

$$e(x) = |\hat{y}(x) - y(x)|.$$  \hfill (3)

A. Uncertainty Estimates for Polynomial Response Surface and Kriging

Polynomial response surface (PRS) approximates the actual function through sum of monomials $f_i(x)$:

$$\hat{y}(x) = \sum_{i=1}^{n_\beta} \beta_i f_i(x),$$  \hfill (4)

The set of $\beta_i$ coefficients can be obtained by least squares according to the PRS theory. Then, the prediction variance at a point, $\varphi_{PRS}^2(x)$, is estimated as:

$$\varphi_{PRS}^2(x) = \left( \frac{(y - \hat{y})^T (y - \hat{y})}{p - n_\beta} \right) f_0^T (F^T F)^{-1} f_0,$$  \hfill (5)

where $y$ and $\hat{y}$ are the values of the actual and estimated responses at the sampled points; $p$ is the number of data points; $n_\beta$ is the number of coefficients of the PRS; and $f_0$ and $F$ are the Gramian matrices (matrices of linear equations constructed using the monomials $f_i(x)$) of the $x$ and the design matrix, respectively. For more details about the PRS prediction variance see Refs. 11,12.

Kriging (KRG) estimates the value of a function as a combination of known functions $f_i(x)$ (e.g., a linear model such as a polynomial trend) and departures (representing low and high frequency variation components, respectively) of the form:

$$\hat{y}(x) = \sum_{i=1}^{m} \beta_i f_i(x) + z(x),$$  \hfill (6)

where $z(x)$ is assumed to be a realization of a stochastic process $Z(x)$ with mean zero, process variance $\sigma^2$, and spatial covariance function given by:

$$\text{cov}(Z(x_i), Z(x_j)) = \sigma^2 R(x_i, x_j),$$  \hfill (7)

$$\sigma^2 = \frac{1}{p} (y - Xb)^T R^{-1} (y - Xb),$$  \hfill (8)

where $R(x_i, x_j)$ is the correlation between $x_i$ and $x_j$; $y$ is value of the actual responses at the sampled points; $X$ is the Gramian design matrix constructed using basis functions in the trend model at the design points; $R$ is the matrix of correlations among design points; and $b$ is an approximation of the vector of coefficients $\beta_i$ of Eq. (6). See Refs. 5,6 for details on how to compute $b$.

We can estimate the uncertainty of $\hat{y}(x)$ using the KRG prediction variance, $\varphi_{KRG}^2(x)$ (also known as mean squared error of the predictor):

$$\varphi_{KRG}^2(x) = \sigma^2 \left( 1 + u^T (X^T R^{-1} X)^{-1} u - r^T X^{-1} u \right),$$  \hfill (9)
where \( u = X^T R^{-1} r - f ; r \) is the vector of correlations between the point \( x \) and the design points; \( f \) is the vector of basis functions in the trend model at point \( x \). For more details about the kriging prediction variance see Refs. 5,6.

### III. Accuracy of Prediction Variance

The accuracy of the prediction variance is measured by its similarity to the prediction error. We use two measures of accuracy. The first is the correlation between the vector of errors, \( e \), and the vector of the square root of the prediction variance, \( \varphi = \sqrt{\varphi^2} \), at a large number of \( p_{test} \) test points:

\[
R = r(e, \varphi). \tag{10}
\]

\( R \) is not an indicator of accuracy of the surrogate model, but an indicator of how well the prediction variance captures the spatial variation of the error (3).

The second measure, \( \rho \), is the ratio between the estimated root mean square error based on the prediction variance, \( \varphi_{RMS} \), and the actual root mean square error, \( e_{RMS} \):

\[
\rho = \frac{\varphi_{RMS}}{e_{RMS}}, \tag{11}
\]

\[
\varphi_{RMS} = \sqrt{\frac{1}{p_{test}} \sum_{i=1}^{p_{test}} \varphi^2_i}, \tag{12}
\]

\[
e_{RMS} = \sqrt{\frac{1}{p_{test}} \sum_{i=1}^{p_{test}} e_i^2} = \sqrt{\frac{1}{p_{test}} \sum_{i=1}^{p_{test}} (\hat{y}_i - y_i)^2}. \tag{13}
\]

Like \( R \), \( \rho \) is also calculated using a large number of \( p_{test} \). A value of \( \rho \) close to 1 indicates good estimate by uncertainty model of overall accuracy. Another way of looking at \( \rho \) is like a scaling factor that makes the \( \varphi_{RMS} \) equals \( e_{RMS} \) (\( \varphi_{RMS}/\rho = e_{RMS} \)). Since the computation of \( R \) and \( \rho \) is not possible from only the training data, we propose to assess the accuracy of the error estimates by cross-validation.

Cross-validation is a process of estimating errors by constructing the surrogate without some of the points and calculating the errors at these left out points. The process is repeated with different sets of left-out points in order to get statistically significant estimates of errors. The process proceeds by dividing the set of \( p \) data points into \( k \) subsets. The surrogate is fit to all subsets except one, and error is checked in the subset that was left out. This process is repeated for all subsets to produce a vector of cross-validation errors, \( e_{XV} \) (also known as the PRESS vector, where PRESS stands for prediction sum of squares). Figure 1-(a) illustrates the cross-validation errors for a kriging surrogate.

Since some surrogates also provide an uncertainty structure, every time that we compute a cross-validation error, we can also compute the associated prediction variance, \( \varphi_{XV} \) (that we called cross-validation error estimate). At the end of the cross-validation process, we would have two vectors, the vector of cross-validation errors, \( e_{XV} \), and the vector of the square root of cross-validation prediction variance, \( \varphi_{XV} \). In our previous work\(^a\), we proposed that the correlation between the error, \( e \), and the square root of the prediction variance, \( \varphi = \sqrt{\varphi^2} \), \( R \) (see Eq. (10)), can be estimated as:

\(^a\) We computed the actual errors using a Latin Hypercube design\(^b\) of 10,000 points created by the MATLAB\(^c\) function \textit{lhsdesign} set with the “\textit{maxmin}” option with 10 iterations.
\[ R_{XV} = r(e_{XV}; \varphi_{XV}) \cdot \]  

The closer \( R_{XV} \) is to one, the more trustful is the uncertainty structure is. Similarly, we can estimate \( \rho \) by:

\[ \rho_{XV} = \frac{\varphi_{RMS_{XV}}}{PRESS_{RMS}}, \]  

\[ \varphi_{RMS_{XV}} = \sqrt{\frac{1}{p} \varphi_{XV}^T \varphi_{XV}}, \]  

\[ PRESS_{RMS} = \sqrt{\frac{1}{p} e_{XV}^T e_{XV}}. \]  

\( PRESS_{RMS} \) is the \( e_{RMS} \) estimator based on \( e_{XV} \). The lower the \( PRESS_{RMS} \) value, the better the overall prediction capability of the surrogate model. We use \( \rho_{XV} \) is the \( \rho \) estimator based on \( \varphi_{XV} \). The closer \( \rho_{XV} \) is to one, the more trustful the uncertainty structure is.

With \( \rho_{XV} \) we scale \( \varphi_{RMS} \) aiming to make the scaled estimated root mean square error based on the prediction variance,

\[ \tilde{\varphi}_{RMS} = \frac{\varphi_{RMS}}{\rho_{XV}}, \]  

a better estimator of \( e_{RMS} \). If the scaled \( \rho \) ratio,

\[ \tilde{\rho} = \frac{\tilde{\varphi}_{RMS}}{e_{RMS}}, \]  

is closer to 1 than the original \( \rho \) is; it means that \( \rho_{XV} \) can also be used to improve the quality of the \( e_{RMS} \) estimation.

Since cross-validation data is available, we also compute the ratio between \( PRESS_{RMS} \) and the actual root mean square error, \( e_{RMS} \):

\[ \rho_{PRESS} = \frac{PRESS_{RMS}}{e_{RMS}}. \]  

### IV. Importing an Error Estimate from a Different Surrogate

#### A. The Idea of Importing an Error Estimate

In some cases, the most accurate surrogate does not provide an error estimate, as shown in Figure 2-(a). Two surrogates, namely kriging (KRG) and support vector regression (SVR) are fitted to the \( \sin(x) \) with five points. The KRG model is set with Gaussian correlation function and a zero order trend function (constant); and the SVR model that uses a Gaussian kernel function and it was fit with the \( \varepsilon \)-insensitive loss function set with \( \varepsilon = 0 \). As a result, both surrogates act as interpolators (they pass through data points). Although SVR is more accurate (smaller \( e_{RMS} \)), KRG would be the choice if we need an error estimate.
We propose is that considering the common points of different surrogate models (interpolation, basis function, etc.), we can combine the predictor of a model with the uncertainty estimate of another model. The idea is illustrated in Figure 2. By comparing Figure 2-(a) and (b), we can see that the kriging offers a reasonable error estimate for the support vector regression.

Another possible scenario is the one where, some reason; one has already decided to use a surrogate that does not provide an error estimate. One way of providing an error estimate is by generating another surrogate that shares some common aspects. In our example, say that we chose the support vector regression. Then, we generate a kriging model to furnish an error estimate.

We might want to check how successfully the imported structure works. If the actual function was known, and therefore the actual errors (Eq. (3)), we could use the same principle of Eqs. (10) and (11) to check how well the errors of surrogate \( a \) are described by the imported prediction variance of surrogate \( b \):

\[
R_{ab} = r\left(e_a, \varphi_b\right),
\]

\[
\rho_{ab} = \frac{\varphi_{RMSb}}{\varepsilon_{RMSa}}.
\]  

The closer \( R_{ab} \) and \( \rho_{ab} \) are to one, the more trustworthy is the imported prediction variance.

When only training data is available, we propose to use cross-validation data to estimate both \( R_{ab} \) and \( \rho_{ab} \). This way:

\[
R_{XVab} = r\left(e_{XV,a}, \varphi_{XV,b}\right),
\]

\[
\rho_{XVab} = \frac{\varphi_{RMS_{XV,b}}}{PRESS_{RMSa}}.
\]

In this case, we can also scale the imported \( \varphi_{RMSb} \):

\[
\tilde{\varphi}_{RMSb} = \frac{\varphi_{RMSb}}{\rho_{XVab}},
\]

and check with a scaled \( \rho_{ab} \) ratio,

\[
\tilde{\rho}_{ab} = \frac{\tilde{\varphi}_{RMSa}}{\varepsilon_{RMSa}},
\]

if cross-validation can also improve the imported \( \varepsilon_{RMS} \) estimator, \( \tilde{\varphi}_{RMSa} \).

**B. Using Multiple Surrogates and Cross-Validation of the Error Estimates**

Since the cost of fitting surrogates is often much smaller than the cost of simulation, it pays to generate a large number of surrogates and select the most accurate based on cross-validation\(^2\). In this situation, if the apparent most accurate surrogate does not have uncertainty estimates, we may be able to choose several candidate uncertainty structures from the subset of surrogates with uncertainty structure. We propose to make the selection based on cross-validation.

First we divide the set of \( n_s \) surrogates into two subgroups, one with \( n_{\varphi} \) surrogates that provide error estimates and the other with the remaining \( n_{\varepsilon} \) surrogates (without error estimates). For all \( n_s \) surrogates, we can use Eq. (23) to compute \( R_{XV} \) between their own vector of cross-validation errors and the vector of cross-validation error...
estimate of different surrogates (obviously belonging to the subgroup that offers error estimates). At the end, we have an $R_{XV}$ matrix ($n_s \times n_\varphi$), in which rows represent different errors and columns represent error estimates:

$$R_{XVij} = r(e_{XVij}, \varphi_{XVj}),$$

$$i = 1, 2, \ldots, n_s$$

$$j = 1, 2, \ldots, n_\varphi$$

(27)

More explicitly, for the $ith$-surrogate of the set (i.e., row of the $R_{XV}$ matrix) the most likely uncertainty estimate will be the one with highest $R_{XVij}$.

Let's consider the example illustrated by Figure 3-(a), where four surrogates, namely a kriging (KRG) model, a polynomial response surface (PRS), a radial basis neural network (RBNN), and a support vector regression (SVR) model are fit to five data points of an arbitrary function, $f(x)$. In our implementation, only KRG and PRS furnish error estimates. Figure 3-(b) shows what we would get from the error estimate of KRG based on the model of Eq. (9). In this case, cross-validation successfully estimates the correlation between error and uncertainty estimate. Figure 3-(c) and (d) illustrate how the error in the RBNN model correlates with the uncertainty estimates of the KRG and PRS surrogates. We can see that (i) it is possible that the actual correlation ($R$) is better than the one of the original surrogate; and (ii) cross-validation successfully selects the best uncertainty estimate to import for the RBNN model. Table 1 shows the complete $R$ and $XV$ matrices for this example. For both KRG and PRS, cross-validation estimates well the actual correlation ($R$). Since PRS approximates while KRG interpolates, there is very poor correlation between one and the other. As for RBNN and SVR, we can see that: (i) we should import the error estimate of KRG to be used with the RBNN predictor; and (ii) although similar rationale would point to the PRS error estimate for the SVR model, given the low values of $R_{XVij}$ and the fact that we use only few (five) data points, we should be aware of a low quality error estimator.

As for the surrogates that already furnish an error estimator (kriging and polynomial response surface), importation is considered only when $R_{XVij} > R_{XVii}$. Thus, for this example, the $R_{XVij}$ values shown in Table 1 do not indicate that we would benefit from switching error estimates for neither kriging nor for polynomial response surface.

V. Numerical Experiments

A. Test Set

We use an analytical example to illustrate each step of the proposed approach. Table 2 details the ten different basic surrogates used during the investigation. The DACE toolbox of Lophaven et al.\textsuperscript{31}, SURROGATES toolbox of Viana\textsuperscript{32}, the native neural networks MATLAB toolbox\textsuperscript{26}, and the code developed by Gunn\textsuperscript{33} were used to execute the kriging, polynomial response surface, radial basis neural network, and support vector regression algorithms, respectively. The SURROGATES toolbox was also used for easy manipulation of the surrogates. No attempt was made to improve the predictions of any surrogate by fine tuning their respective parameters (such as the initial correlation parameter in kriging models). In this work we use multiple instances of different surrogates (in the same fashion of Refs. 22 and 34).

As test problems, we employed the two following widely used analytical benchmark problems\textsuperscript{35}:

- Branin-Hoo (2 variables):

$$y(x) = \left\{ x_2 - \frac{5.1x_1^2} {4\pi^2} + \frac{5x_1} {\pi} - 6 \right\}^2 + 10\left\{1 - \frac{1} {8\pi}\right\} \cos(x_1) + 10,$$

$$-5 \leq x_1 \leq 10, \quad 0 \leq x_2 \leq 15.$$  

(28)
• Hartman6 (6 variables):

\[ y(x) = -\sum_{i=1}^{4} a_i \exp \left( -\sum_{j=1}^{6} B_{ij} (x_j - D_{ij})^2 \right), \]

\[ 0 \leq x_j \leq 1, \ j = 1, 2, \ldots, 6, \]

\[ B = \begin{bmatrix} 10.0 & 3.0 & 17.0 & 3.5 & 1.7 & 8.0 \\ 0.05 & 10.0 & 17.0 & 0.1 & 8.0 & 14.0 \\ 3.0 & 3.5 & 1.7 & 10.0 & 17.0 & 8.0 \\ 17.0 & 8.0 & 0.05 & 10.0 & 0.1 & 14.0 \end{bmatrix}, \]

\[ D = \begin{bmatrix} 0.1312 & 0.1696 & 0.5569 & 0.0124 & 0.8283 & 0.5886 \\ 0.2329 & 0.4135 & 0.8307 & 0.3736 & 0.1004 & 0.9991 \\ 0.2348 & 0.1451 & 0.3522 & 0.2883 & 0.3047 & 0.6650 \\ 0.4047 & 0.8828 & 0.8732 & 0.5743 & 0.1091 & 0.0381 \end{bmatrix}. \]

To investigate the effect of the point density, we fitted the Branin-Hoo function using both 20 and 42 points and the Hartman6 function with 56 and 110 points.

The quality of fit, and thus the performance, depends on the design of experiment (DOE). Hence, for all test problems, a set of 1000 different Latin Hypercube designs were used as a way of averaging out the DOE dependence of the results. We used the MATLAB function `lhsdesign`, set with the “maximin” option with 1000 iterations to generate the DOEs for fitting.

B. Results and Discussion

Among the ten surrogates presented in Table 2, just five provide prediction variance. We first investigate how good the original prediction variance is as an uncertainty structure. Figure 4 illustrates the boxplots of the correlation, \( R \), between the error and the square root of the prediction variance at test points. See Eq. (10) for information about the computation of \( R \). Appendix A details boxplots. Data is gathered out of 1000 design of experiments (DOEs) for surrogates that provide prediction variance. We see that the correlation can be substantially smaller than 1, meaning that the prediction variance does not always describe well the actual errors. As expected, the sparseness of the DOE in the six-dimensional space negatively affects both the kriging and polynomial response surface prediction variances. However, the cubic polynomial response surface can surprisingly offer better uncertainty structure than kriging for 110 points. As discussed in our previous work\(^\text{22}\), the contribution of the correlation function in kriging fades with the curse of dimensionality. Then, kriging models are more influenced by the trend function. In terms of the prediction variance, regions that are close to the data points benefit from the interpolation and correlation function, but they represent just a small portion of the design space. The result is that the prediction variance tends to the kriging process variance. This explains the degeneration of the kriging performance when moving from Branin-Hoo to Hartman6. It also explains comparable performances of KRG2 and PRS2 for Hartman6 with 56 points. With 110 points, Both KRG2 and PRS2 have performance degraded, indicating improper order of the trend or polynomial. Another important observation is that the uncertainty estimates of the PRS models suffer from adding points. The prediction variance of the PRS assumes that the data has errors, and usually the PRS is more accurate than the data at the points of the design of experiment. Increasing the number of points makes the prediction variance smaller compared to the error at data points (DOE). So at data points, the prediction variance can be very small even though the error is large (away from data points the discrepancy is not so pronounced). With the Branin-Hoo function this effect is much more evident than with the Hartman6.

Figure 5 complements Figure 4 showing the boxplots of the ratio between the estimated root mean square error based on the prediction variance, \( \rho \), its scaled version, \( \tilde{\rho} \), and the version obtained with \( \text{PRESS}_{\text{RMS}} \cdot \rho_{\text{PRESS}} \). Eq. (11), (19), and (20) detail the computation of the ratios. Without scaling (i.e., considering \( \rho \)), there is a common tendency in underestimating \( e_{\text{RMS}} \) (since \( \rho \) tends to be smaller than 1). However, the differences between kriging and the polynomial response surface models are less pronounced with the increase in variables. Note that for Branin-Hoo, the third order polynomial response surface is less efficient than the second order polynomial response surface. This is because we are not considering the accuracy (\( e_{\text{RMS}} \), see Table 3) but the point-wise uncertainty estimation. The prediction variance of the PRS assumes that the data has errors, and usually the PRS is more accurate than the data. If the number of points is very large, the prediction variance becomes small compared to the difference between the data and the PRS. Consequently, the prediction variance can be very small at data points even though the error is large (this is not so pronounced away from data points). We can see that scaling (i.e. considering \( \tilde{\rho} \)
improves the $e_{RMS}$; because $\hat{\rho}$ is generally closer to 1 than $\rho$ is. This means that $\rho_{XV}$ satisfactorily estimates $\rho$. In fact, the results are comparable with those obtained with $PRESS_{RMS}$. The only problem appears for all kriging models of Branin-Hoo fitted with 42 points. Counter-intuitively, adding more points made the estimation less accurate. For illustration, we show in Table 4 the statistics for the $e_{RMS}$ estimation for the kriging model KRG0. We can clearly see cross-validation does not offer significant data neither for $e_{RMS}$ estimation nor $\varphi_{RMS}$ correction. (i.e., cross-validation gives a poor correction factor $\rho_{XV}$).

Next, we investigate how our proposed importation succeeds in providing uncertainty estimation for surrogates. Figure 6 illustrates how well the kriging prediction variance can characterize the errors in kriging and support vector regression models. Figure 6-(a) repeats the boxplots of $R$ from Figure 4-(c) for two kriging models. This is done for easy comparison with other parts of the figure. Figure 6-(b) illustrates the performance when a support vector regression surrogate imports the prediction variance from those kriging models. The $R$ values are actually better in the support vector regression model. It means that the prediction variance can capture the trends of the actual error. It is clear, however, that the performance is dependent on which surrogate the uncertainty is imported from. Figure 7 completes the analysis of the importation of uncertainty structures showing different ratios between the estimated and the actual root mean square error. We arbitrarily chose one of the kriging models for illustration. In this case, we can see that scaling adapts the imported uncertainty structure to better estimate $e_{RMS}$.

The difficulty in predicting which surrogate will perform better motivates the use of multiple surrogates and cross-validation. We first investigate how well cross-validation can point to the best combination of predictor and uncertainty structure. In a design of experiment (DOE) we test that by building both the matrix of correlation between the error and the square root of the prediction variance at a large number of test points, i.e. the $R$ matrix defined in Eq. (21); and the estimate of the $R$ matrix based on cross-validation, i.e. the $R_{XV}$ matrix defined in Eq. (23). Cross-validation succeeds when the selection based on the $R$ and $R_{XV}$ matrices leads to same/similar performance (what is important is the final correlation that the imported uncertainty structure has with the error of the surrogate). Figure 8 illustrates the boxplots out of 1000 DOEs of the $R$ and $\rho$ (see Eq. (11)) for the surrogates with uncertainty estimates and surrogates selected from the $R$ and $R_{XV}$ matrices for two test problems. Similar observations were found for the other cases. The surrogate selected from the $R$ matrix represents the solution of choice if we knew the performance beforehand. While not practical, it is the maximum achievable. Again, the performance of individual surrogates changes with the problem, but we can see that the surrogate selected from the $R_{XV}$ matrix offers an excellent compromise solution, meaning that it is closer to the surrogate selected from the $R$ matrix. We further checked which of the choices pointed by the $R$ and $R_{XV}$ matrices represent importation of off-diagonal elements of the matrices. Table 5 shows the percentage of the cases that happens. As expected, increasing the number of points makes the uncertainty structures better, increasing the frequency in which the highest values are diagonal elements.

Next, we study how well cross-validation point to the surrogate which the uncertainty structure should be imported from. Figure 9-(a) shows boxplots of the $R$ values for the surrogates with uncertainty estimates and surrogates which uncertainty is imported based on the $R_{XV}$ matrix (correlation taken with cross-validation data). Similar results were observed in the other cases. Considering that the imported uncertainties are not based on the statistical assumptions of the surrogates, we can say that cross-validation succeeds in points the surrogates for uncertainty importation. Figure 9-(b) shows boxplots of the $R$ values for the one of the support vector regression when the uncertainty is imported based both on the $R_{XV}$ and the $R$ matrices ($R$ means correlation taken with test points). The selection based on cross-validation is comparable with the selection based on test data. Again, the estimation of $e_{RMS}$ completes the study. Figure 10 gives the boxplots of $\hat{\rho}_{ab}$ and $\rho_{PRESS}$, see Eq. (26) for details. As expected, point density improves the accuracy of the estimation. But surprisingly, again, the scaled imported uncertainty structure provides a comparably accurate and less noisy estimate of $e_{RMS}$ compared to $PRESS_{RMS}$.

Last but not least, we want to see if choosing the uncertainty structure based on the overall error (i.e., minimum $e_{RMS}$ by using $PRESS^{24}$) also works. Table 6 shows the frequency in which the surrogates of Table 2 selected for
reduced $e_{\text{RMS}}$ or $PRESS_{\text{RMS}}$ do not furnish an uncertainty structure. Not surprisingly, the best surrogate of choice may not furnish an uncertainty structure. Considering the subset of surrogates with uncertainty structure, Figure 11 shows the median of the $R$ values (out of 1000 DOEs) for surrogates picked according to different criteria (including both prediction and uncertainty structure). Unexpectedly, the surrogates chosen based on the prediction (using either $e_{\text{RMS}}$ or $PRESS_{\text{RMS}}$) offer a rather limited performance in terms of their uncertainty structure. Although counter-intuitive, this leads to the conclusion that even though the overall prediction is better, the uncertainty estimation may not be.

VI. Conclusion

We proposed two schemes for importation of uncertainty estimates. One is the arbitrary importation of uncertainty structure from another surrogate. The other one is based on the generation of multiple surrogates and cross-validation for uncertainty structure selection. We used two algebraic examples and ten basic surrogates (including different instances of kriging, polynomial response surface, radial basis neural networks and support vector regression surrogates) as test cases in our study. For these examples we found that

- The statistically based uncertainty estimates may not always correlate well to the errors. As a consequence, there is a risk that the surrogate chosen based on prediction capabilities (minimum root mean square error) will provide poor uncertainty estimation.
- Imported uncertainty structure performed reasonably well when compared to the statistically-based ones.
- Scaling the imported uncertainty improved the estimation of $e_{\text{RMS}}$ to comparable accuracy with estimation based on $PRESS_{\text{RMS}}$. We hope that with that we can also improve the prediction variance as a point-wise uncertainty measure (by using $\rho_{XY}$ as a scaling factor). This would enhance the performance of algorithms like the Efficient Global Optimization (EGO) algorithm.

And about the use of cross-validation, we conclude that:

- Cross-validation of the prediction variance effectively ranked the surrogates according to the quality of the uncertainty structure. We saw that this allowed not only defining which surrogate we should import the uncertainty from, but also to point to the best surrogate of the set.
- The amplitude correction factor improved amplitude of the uncertainty structure based on the estimation of the root mean square error based on the integration of the prediction variance.

Appendix

A. Boxplots

In a boxplot, the box is defined by lines at the lower quartile (25%), median (50%), and upper quartile (75%) values. Lines extend from each end of the box to show the coverage of the rest of the data. Outliers are data with values beyond the ends of the lines by placing a "+" sign for each point.

Acknowledgments

This work has been supported by the NASA Constellation University Institute Program (CUIP) and the National Science Foundation (Grant # 0423280). Authors gratefully acknowledge these grants.

References


Table 1: Correlation between error and square root of prediction variance, $R_{ij} = r(e_i, \varphi_j)$, and its cross-validation estimate, $R_{X_{ij}} = r(e_{X_{ij}}, \varphi_{X_{ij}})$, of the example illustrated by Figure 3.

<table>
<thead>
<tr>
<th>Surrogates</th>
<th>KRG</th>
<th>PRS</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>$R_{XY}$</td>
<td>R</td>
</tr>
<tr>
<td>KRG</td>
<td>0.6</td>
<td>0.6</td>
</tr>
<tr>
<td>PRS</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>RBNN</td>
<td>0.8</td>
<td>0.7</td>
</tr>
<tr>
<td>SVR</td>
<td>0.4</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Table 2: Information about the set of 10 surrogates. In the current implementation, just 5 of them provide uncertainty estimates.

<table>
<thead>
<tr>
<th>Surrogates</th>
<th>Details</th>
<th>Provide uncertainty estimate?</th>
</tr>
</thead>
<tbody>
<tr>
<td>krg0</td>
<td>Kriging models: KRG0, KRG1, and KRG2 indicate zero, first, and second order polynomial regression model, respectively. In all cases, a Gaussian correlation and $\theta_{0i} = { -\frac{1}{n_v} }$, and $10^{-3} \leq \theta_i \leq 2 \times \theta_{0i}, i = 1,2,\ldots,n_v$ were used. We chose 3 different kriging surrogates by varying the regression model.</td>
<td>YES</td>
</tr>
<tr>
<td>krg1</td>
<td>Radial basis neural network: $Goal = (0.05\overline{y})^2$ and $Spread = \frac{1}{\sqrt[3]{3}}$.</td>
<td>NO</td>
</tr>
<tr>
<td>krg2</td>
<td>Support vector regression: GRBF and Poly indicate the kernel function (Gaussian and second order polynomial respectively).</td>
<td></td>
</tr>
<tr>
<td>rbnn</td>
<td>All use loss function as $\varepsilon$-insensitive and quadratic, respectively. Full and Short refer to different values for the regularization parameter, $C$, and for the insensitivity, $\varepsilon$. Full adopts $C = \infty$ and $\varepsilon = 1 \times 10^{-4}$, while Short uses the selection of values according to Cherkassky and Ma$^{30}$: $\varepsilon = \sigma_{\bar{y}} / \sqrt{p}$; and for both $C = \max \left( \left</td>
<td>\overline{y} + 3\sigma_{\overline{y}} \right</td>
</tr>
<tr>
<td>svr-grbf-full</td>
<td>We chose 4 different SVR surrogates by varying the kernel function and the SVR parameters ($C$ and $\varepsilon$).</td>
<td></td>
</tr>
<tr>
<td>svr-grbf-short</td>
<td></td>
<td></td>
</tr>
<tr>
<td>svr-poly-full</td>
<td>Polynomial response surface: Full models of degree 2 and 3.</td>
<td>YES</td>
</tr>
<tr>
<td>svr-poly-short</td>
<td></td>
<td></td>
</tr>
<tr>
<td>prs2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>prs3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Statistics (out of 1000 DOE samples) about the $\varepsilon_{RMS}$ of PRS2 and PRS3 for the Branin-Hoo function. See Table 2 for details about the surrogate models.

<table>
<thead>
<tr>
<th>Surrogate</th>
<th>Number of points</th>
<th>20</th>
<th>42</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>median</td>
<td>standard deviation</td>
<td>median</td>
</tr>
<tr>
<td>PRS2</td>
<td>28.5</td>
<td>2.5</td>
<td>26.9</td>
</tr>
<tr>
<td>PRS3</td>
<td>8.3</td>
<td>1.8</td>
<td>7.0</td>
</tr>
</tbody>
</table>
Table 4: Statistics (out of 1000 DOEs) about the $e_{RMS}$ estimation of the kriging model, KRG0, fitted with 42 points for the Branin-Hoo function ($\rho = \varphi_{RMS}/e_{RMS}$, $\bar{\rho} = \varphi_{RMS}/e_{RMS}$, and $\rho_{PRESS} = PRESS_{RMS}/e_{RMS}$). See Table 2 for details about KRG0. Cross-validation does not generate meaningful data for correction of $\varphi_{RMS}$.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>$\rho$</th>
<th>$\bar{\rho}$</th>
<th>$\rho_{PRESS}$</th>
<th>$e_{RMS}$</th>
<th>$\varphi_{RMS}$</th>
<th>$\bar{\varphi}_{RMS}$</th>
<th>PRESS$_{RMS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>0.7</td>
<td>1.5</td>
<td>3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.5</td>
<td>8.6</td>
</tr>
<tr>
<td>Std deviation</td>
<td>0.7</td>
<td>3</td>
<td>110</td>
<td>6</td>
<td>3</td>
<td>6</td>
<td>31</td>
</tr>
</tbody>
</table>

Table 5: % Frequency (out of 1000 DOEs) in which the highest term of the $R$ and $R_{XY}$ matrices are in the diagonal. See Eqs. (21) and (23) for details about the matrices. Being in the diagonal means that there is no need for importation of uncertainty structure.

<table>
<thead>
<tr>
<th>Matrix</th>
<th>Branin-Hoo</th>
<th>Hartman6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20 points</td>
<td>42 points</td>
</tr>
<tr>
<td>$R$</td>
<td>32.9</td>
<td>70.4</td>
</tr>
<tr>
<td>$R_{XY}$</td>
<td>34.4</td>
<td>70</td>
</tr>
</tbody>
</table>

Table 6: % Frequency (out of 1000 DOEs) in which the best surrogate of the set (given in Table 2) according to the prediction capabilities does not furnish an uncertainty structure. $e_{RMS}$ allows the actual best surrogate. PRESS$_{RMS}$ gives the selection based on data points.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Branin-Hoo</th>
<th>Hartman6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20 points</td>
<td>42 points</td>
</tr>
<tr>
<td>$e_{RMS}$</td>
<td>0.4</td>
<td>25.1</td>
</tr>
<tr>
<td>PRESS$_{RMS}$</td>
<td>1.8</td>
<td>20.8</td>
</tr>
</tbody>
</table>
Figure 1 Cross-validation measurements exemplified by fitting a kriging model (KRG) to \( p = 5 \) data points of the function \( \sin(x) \). Adopted from Ref. 24. Gray areas illustrate the square root of the prediction variance. (a) Cross-validation error at the fourth point of the DOE, \( e_{XV_4} \); and (b) Cross-validation error, \( e_{XV_4} \), and the square root of the cross-validation prediction variance, \( \phi_{XV_4} \) (details for the fourth point of the DOE).

Figure 2: Importation of uncertainty estimates. Figure 2-(a) shows a kriging (KRG) model set with Gaussian correlation function and a support vector regression (SVR) model that uses a Gaussian kernel function and was fit with the \( \varepsilon \)-insensitive loss function set with \( \varepsilon = 0 \) (both surrogates act as interpolators). Although KRG is less accurate, it would be the choice if we need an uncertainty estimate. Figure 2-(a) shows the KRG model and the square root of its prediction variance. Figure 2-(b) shows the SVR model with the square root of the prediction variance borrowed from KRG. Although not based on the SVR statistical assumptions, the KRG uncertainty estimate offers a reasonable solution.
Figure 3: Using multiple surrogates and cross-validation for importation of uncertainty estimates. Figure 3-(a) shows the fit of four different surrogates, namely a kriging (KRG) model, a polynomial response surface (PRS), a radial basis neural network (RBNN), and a support vector regression (SVR) model, to an arbitrary set of five points. Among these surrogates, only the KRG and the PRS models provide an uncertainty estimate. This is exemplified by the KRG model and the square root of its prediction variance in Figure 3-(b). Figure 3-(c) and (d) illustrates the importation of different uncertainty estimates for the RBNN model. As shown in Table 1, using cross-validation we would select for the RBNN model the uncertainty estimated imported from KRG.

Figure 4: Boxplots of the correlation between error and uncertainty estimates, \( R = R(e, \varphi) \), for surrogates that offer an error estimate (out of 1000 DOEs). See Eq. (10) for details about computation of \( R \). Table 2 describes each surrogate model. It can be seen that the increased dimensionality affects kriging models more than polynomial response surface ones. Appendix A details boxplots.
Figure 5: Boxplots of the ratio between different estimators of the root mean square error ($\hat{\varphi}_{\text{RMS}}$, $\tilde{\varphi}_{\text{RMS}}$, and $\text{PRESS}_{\text{RMS}}$), and the actual root mean square error, $e_{\text{RMS}}$, for surrogates that offer an error estimate (out of 1000 DOEs). See Eqs. (11), (19), and (20) for details about computation of the ratios. Table 2 describes each surrogate model. Appendix A details boxplots. We see that (i) the more points the better $\hat{\varphi}_{\text{RMS}}$ estimates $e_{\text{RMS}}$; and (ii) the correction is most likely to work.
(a) Performance of two of the kriging uncertainty structure.

(b) Importing uncertainty structures for svr-grbf-full from different kriging models.

Figure 6: Boxplots of the correlation between error and square root of the prediction variance, $R = r(e, \varphi)$ (out of 1000 DOEs) for Hartman6 with 56 points. See Table 2 for details about each surrogate model. Figure 6-(a) shows the performance of two of the kriging models (they should be seen as references). Figure 6-(b) illustrates that the performance of the imported uncertainty estimate (i) depends on which surrogate the structure is imported from; and (ii) may be comparable to what happens with the original surrogate. Appendix A details boxplots.

Figure 7: Boxplots of the ratio between the estimated root mean square error based on the prediction variance and the actual root mean square error (out of 1000 DOEs) for Hartman6 with 56 points. See Table 2 for details about each surrogate model. See Eqs (19), (22), and (26) for details about $\tilde{\rho}$, $\rho_{ab}$, and $\tilde{\rho}_{ab}$, respectively. Figure 6 $\tilde{\rho}$ of KRG0 should be seen as reference (statistical-based estimation of the $e_{RMS}$). We can see that it is important to scale the imported uncertainty estimate. Appendix A details boxplots.
Figure 8: Performance of different surrogates in terms of uncertainty structure. The kriging and polynomial response surface models used original uncertainty estimates. The surrogate combines the predictor of one of the 10 surrogates of Table 2 and the prediction variance picked according to the Rxv matrix. Eq. (27) details the Rxv matrix. Cross-validation successfully selects the surrogate according to the quality of the uncertainty structure. Appendix A details boxplots.

Figure 9: Cross-validation for importation of uncertainty structure. Boxplot presents the data for Branin-Hoo, fitted with 20 points (out of 1000 DOE’s). Data of original error and (formulation-based) prediction variance of surrogates that furnish uncertainty estimators is shown in gray. In Figure 9-(a), black shows when prediction variance imported and gray when it is not. In Figure 9-(b), black shows when prediction variance imported using cross-validation data and gray when it uses test points. Cross-validation delivers imported uncertainty structures that are comparable not only with statistical-based ones, but also with the ones that would be chosen based on test data. Appendix A details boxplots.
Figure 10: Boxplots of the ratio between different estimators of the root mean square error ($\frac{\hat{\phi}_{RMSb}}{\hat{\phi}_t}$ and $\frac{RMS_{PRESS}}{RMS_e}$), and the actual root mean square error, $c_{RMS}$, for surrogates that offer an error estimate (out of 1000 DOEs). The closer the ratios are to 1, the better. See Eqs. (20) and (26) for details about computation of the ratios. Table 2 describes each surrogate model. Appendix A details boxplots. We see that $\hat{\phi}_{RMSb}$ offers an estimator as good as $PRESS_{RMS}$ does.
Figure 11: Median (out of 1000 DOEs) of the correlation between error and original uncertainty estimates, $R = r(e, \varphi)$ of surrogates selected according to different criteria. From all surrogates of Table 2, we just consider the subset of that furnish uncertainty estimates. Surrogates selected according to $e_{RMS}$ and $PRESS_{RMS}$ may not have an uncertainty structure; see Table 6. The best surrogate in terms of prediction (selected according to $e_{RMS}$) may not be the best surrogate in terms of uncertainty structure. Cross-validation successfully selects the surrogate according to the uncertainty structure.