Complex Computer Models: Leveraging Methods for Physical Experiments

G. Alexi Rodríguez Arelis and William J. Welch

University of British Columbia

Latest Advances in the Theory and Applications of
Design and Analysis of Experiments
BIRS
August 7, 2017

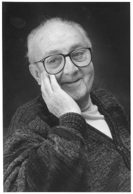
Table of contents

Physical Experiments: Bad Boys

- High dimensionality of factors
- Nonlinear relationships (factors with many levels)
- High-order interaction effects
- Substantial random variation

Physical Experiments: Good Guys (Statisticians)

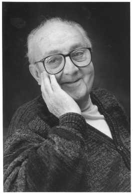
- Effect sparsity (Box and Meyer, 1985).
  - Only some factors are active
  - Only some (low-order) effects are important



DavidMCEddy at en.wikipedia

Physical Experiments: Good Guys (Statisticians)

- Effect sparsity (Box and Meyer, 1985).
  - Only some factors are active
  - Only some (low-order) effects are important



DavidMCEddy at en.wikipedia

- Effect heredity:
  significant interaction $\Rightarrow$ at least one of its factors has a significant main effects (Yates, 1937; Hamada and Wu, 1992)
- Effect hierarchy: lower-order effects tend to be more significant than higher-order effects (Yates, 1937; Wu and Hamada, 2009)

Physical Experiments: Tools

- Many, many other contributors . . .
- Main effects plans
- Resolution IV and V designs give up up on estimation of 3-factor and higher-order interaction effects
- etc., etc.

Physical Experiments: Tools

- Many, many other contributors . . .
- Main effects plans
- Resolution IV and V designs give up up on estimation of 3-factor and higher-order interaction effects
- etc., etc.
- What is the point here for computer experiments?

Computer Experiments: Bad Boys

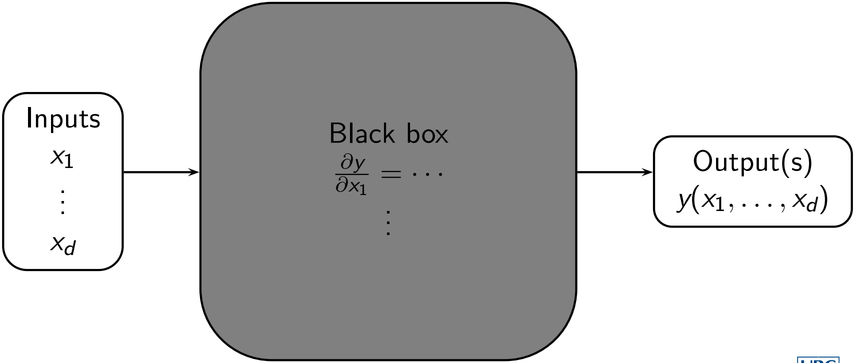
- High dimensionality of inputs
- Nonlinear relationships
- High-order interaction effects
- ~~Substantial random variation~~
- WHATCHA GONNA DO?

Computer Experiments: Bad Boys

- High dimensionality of inputs
- Nonlinear relationships
- High-order interaction effects
- WHATCHA GONNA DO?
  - You can cope with 2 bad boys but not 3
  - Standard GP approaches try to deal with all 3 simultaneously

Computer Model

- Inputs $\mathbf{x} = (x_1, \ldots, x_d)$
- Output $y = y(\mathbf{x})$



Gaussian Process (GP) Model

- Model $y(\mathbf{x})$ as a realization of

$$Y(\mathbf{x}) = \mu + Z(\mathbf{x})$$

- μ is a constant mean; regression model rarely helpful (Chen et al., 2016)
- $Z(\mathbf{x})$ is a multivariate normal random function with
  - Marginal distribution $\mathrm{N}\left(0, \sigma^2\right)$
  - Covariance $\mathrm{Cov}(Z(\mathbf{x}), Z(\mathbf{x}')) = \sigma^2 R(\mathbf{x}, \mathbf{x}')$ at two input vectors $\mathbf{x}$ and $\mathbf{x}'$

Gaussian Process (GP) Model

- Model $y(\mathbf{x})$ as a realization of

$$Y(\mathbf{x}) = \mu + Z(\mathbf{x})$$

- μ is a constant mean; regression model rarely helpful (Chen et al., 2016)
- $Z(\mathbf{x})$ is a multivariate normal random function with
  - Marginal distribution $\mathrm{N}\left(0, \sigma^2\right)$
  - Covariance $\mathrm{Cov}(Z(\mathbf{x}), Z(\mathbf{x}')) = \sigma^2 R(\mathbf{x}, \mathbf{x}')$ at two input vectors $\mathbf{x}$ and $\mathbf{x}'$
  - Correlation function $R(\mathbf{x}, \mathbf{x}')$ is critical

Correlation Function

- Two sets of inputs $\mathbf{x}$ and $\mathbf{x}'$
- Standard (Std) correlation structure has a product form in 1-dimensional functions

$$R^{\mathrm{Std}}(\mathbf{x}, \mathbf{x}') = \prod_{j=1}^{d} R(h_j) \quad \in [0, 1],$$

  a function of the distance $h_j = |x_j - x_j'|$ for input j
- We will use the squared-exponential (Gaussian) function

$$R(h_j) = \exp(-\theta_j h_j^2),$$

  where $\theta_j \geq 0$ controls the sensitivity (nonlinearity) of input j.
- Following arguments apply to correlation functions other than squared exponential

Estimation

- μ, σ^2, and the correlation parameters $(\theta_1, \ldots, \theta_d)$ estimated by MLE or Bayes
- For the results shown later, $\theta_1, \ldots, \theta_d$ optimized numerically via the `mlegp` package (Dancik, 2013)
- All standard stuff
- (`mlegp` is adapted for the new correlation structures introduced later)

Prediction

- Predict $y(\mathbf{x}^*)$ at some new input vector $\mathbf{x}^*$
- More standard stuff ...
- The best linear unbiased predictor or posterior mean (given the correlation parameters) is of the form

$$\hat{y}(\mathbf{x}^*) = \hat{\mu} + \mathbf{a}^T \mathbf{r}(\mathbf{x}^*)$$

  - the vector $\mathbf{a}$ does not depend on $\mathbf{x}^*$
  - $\mathbf{r}(\mathbf{x}^*) = \left(R(\mathbf{x}^*, \mathbf{x}^{(1)}), \dots, R(\mathbf{x}^*, \mathbf{x}^{(n)})\right)^T$ is a vector of correlations between $Y(\mathbf{x}^*)$ and $Y(\mathbf{x}^{(i)})$ at the training design points $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$

Prediction

- Predict $y(\mathbf{x}^*)$ at some new input vector $\mathbf{x}^*$
- More standard stuff ...
- The best linear unbiased predictor or posterior mean (given the correlation parameters) is of the form

$$\hat{y}(\mathbf{x}^*) = \hat{\mu} + \mathbf{a}^T \mathbf{r}(\mathbf{x}^*)$$

  - the vector $\mathbf{a}$ does not depend on $\mathbf{x}^*$
  - $\mathbf{r}(\mathbf{x}^*) = \left(R(\mathbf{x}^*, \mathbf{x}^{(1)}), \ldots, R(\mathbf{x}^*, \mathbf{x}^{(n)})\right)^T$ is a vector of correlations between $Y(\mathbf{x}^*)$ and $Y(\mathbf{x}^{(i)})$ at the training design points $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$
  - The predictor has the same form for any legal correlation function
  - The predictor is a linear combination of basis functions $R(\mathbf{x}^*, \mathbf{x}^{(i)})$

Prediction Accuracy of $\hat{y}$

- Measured by root mean squared error (RMSE) over a new test set
- Normalized version compares with the RMSE of the trivial predictor $\bar{y}$, the sample mean of the training data:

$$e_{\text{N-RMSE}} = \frac{\text{RMSE of } \hat{y}}{\text{RMSE of } \bar{y}}$$

- $e_{\text{N-RMSE}} = 0$ says perfect accuracy
- $e_{\text{N-RMSE}} = 1$ says $\hat{y}$ no better than $\bar{y}$

Additive Correlation Structures

- Recall

$$R^{\mathrm{Std}}(\mathbf{x}, \mathbf{x}') = \prod_{j=1}^{d} \exp(-\theta_j h_j^2) \quad (\text{where } h_j = x_j - x_j')$$

- Instead, additive correlation structure with equal weights (A-EW)

$$R^{\mathrm{A-EW}}(\mathbf{x}, \mathbf{x}') = \frac{1}{d} \sum_{j=1}^{d} \exp(-\theta_j h_j^2)$$

- Or additive correlation structure with unequal weights (A-UW)

$$R^{\mathrm{A-UW}}(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{d} \omega_j \exp(-\theta_j h_j^2)$$

  - weights $\omega_1, \ldots, \omega_d$ have to be estimated too, subject to $\omega_j \geq 0$ and $\sum_{j=1}^{d} \omega_j = 1$

Michalewicz Test Function (Ba and Joseph, 2012)

- Michalewicz function in d dimensions:

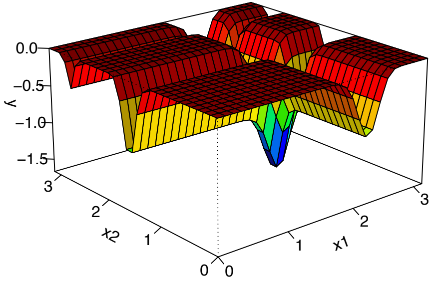
$$y_{\mathrm{u}}(\mathbf{x}) = -\sum_{j=1}^{d} \sin(x_j)\left[\sin\left(\frac{jx_j^2}{\pi}\right)\right]^{20} \quad \text{where} \quad x_j \in [0, \pi].$$

- Has $d!$ local minima (highly nonlinear)
- Weighted form:

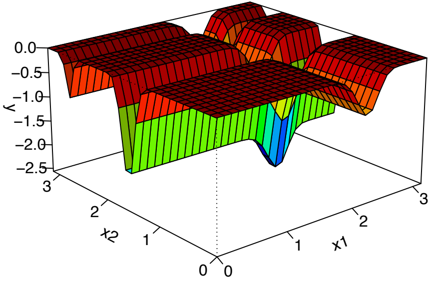
$$y_{\mathrm{w}}(\mathbf{x}) = -\sum_{j=1}^{d} j \sin(x_j)\left[\sin\left(\frac{jx_j^2}{\pi}\right)\right]^{20} \quad \text{where} \quad x_j \in [0, \pi].$$

  - Complexity and more importance of x_j increases with j

Michalewicz Test Function

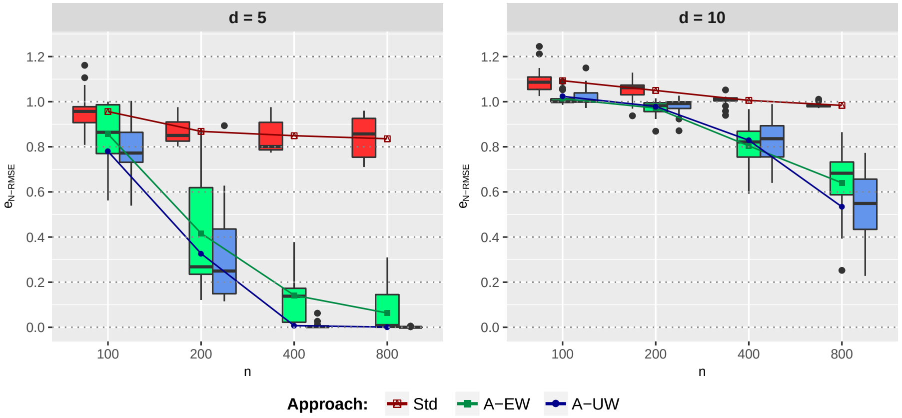


(a) Unweighted



(b) Weighted

Weighted Michalewicz Function ($e_{\text{N-RMSE}}$ for 20 Repeat Training Experiments)



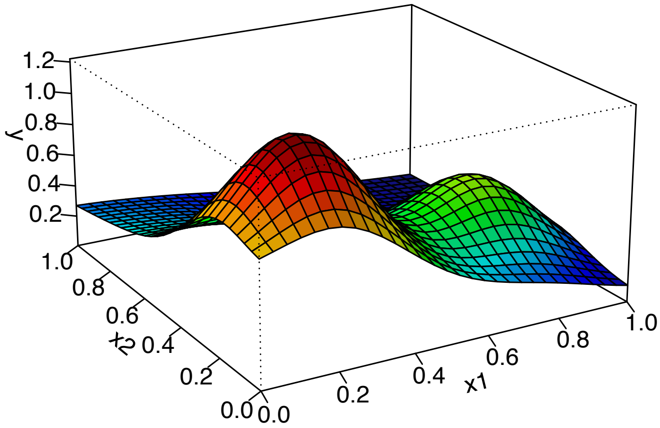
Franke's Function (Franke, 1979; Haaland and Qian, 2011)

- Most functions are not perfectly additive
- Franke's function in 2 dimensions:

$$
\begin{aligned}
y_{\mathrm{F}}(x_1, x_2) = 0.75 \exp\left[-\frac{(9x_1-2)^2}{4} - \frac{(9x_2-2)^2}{4}\right] \\
+ 0.75 \exp\left[-\frac{(9x_1+1)^2}{49} - \frac{9x_2+1}{10}\right] \\
+ 0.5 \exp\left[-\frac{(9x_1-7)^2}{4} - \frac{(9x_2-3)^2}{4}\right] \\
- 0.2 \exp[-(9x_1-4)^2 - (9x_2-7)^2] \\
\text{where} \quad x_1, x_2 \in [0, 1].
\end{aligned}
$$

- Nonlinear with interaction

Franke's Function in 2 Dimensions



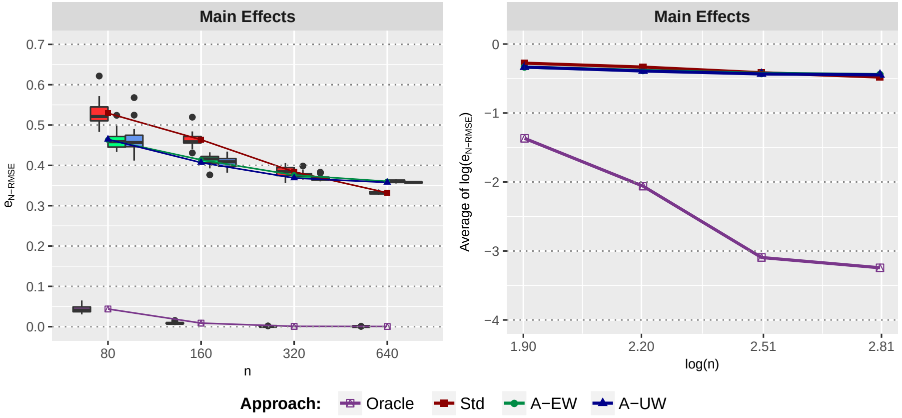
New Franke's Function in 8 Dimensions

- $y_{\mathrm{F8}}(\mathbf{x}) = y_{\mathrm{F}}(x_1, x_2) + y_{\mathrm{F}}(x_3, x_4) + y_{\mathrm{F}}(x_5, x_6) + y_{\mathrm{F}}(x_7, x_8)$
- Nonlinear with moderate interaction (x_1 with x_2, x_3 with x_4, etc.)

Oracle

- Just for a best possible baseline; cannot be done in practice
- Knows the structure of $y_{\mathrm{F8}}(\mathbf{x})$ and fits 4 separate GPs
  - Training data for x_1, x_2, $y_{\mathrm{F}}(x_1, x_2)$ gives $\hat{y}_{\mathrm{F},12}(x_1, x_2)$
  - $\cdots$
  - Training data for x_7, x_8, $y_{\mathrm{F}}(x_7, x_8)$ gives $\hat{y}_{\mathrm{F},78}(x_7, x_8)$
- $\hat{y}_{\mathrm{F8}}^{(\mathrm{Oracle})}(\mathbf{x}^*) = \hat{y}_{\mathrm{F},12}(x_1^*, x_2^*) + \cdots + \hat{y}_{\mathrm{F},78}(x_7^*, x_8^*)$

Franke's Function in 8 Dimensions: Additive (Main Effects) Model



Normalized RMSE versus *n* and log-log plot to show rate of convergence

Composite Correlation Structures (Additive and All 2-Input Interactions)

- Composite, equal weights

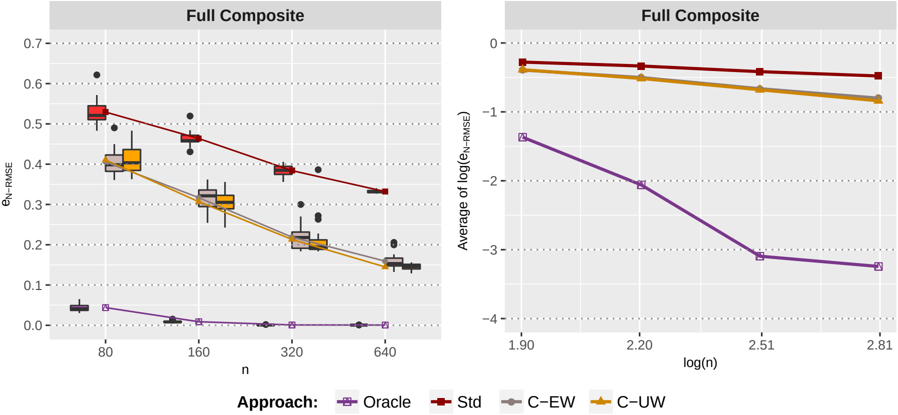
$$R^{\mathrm{C-EW}}(\mathbf{x}, \mathbf{x}') = \frac{1}{2} \cdot \frac{1}{d} \sum_{j=1}^{d} e^{-\theta_j h_j^2} + \frac{1}{2} \cdot \frac{2}{d(d-1)} \sum_{j=1}^{d} \sum_{j'=j+1}^{d} e^{-\theta_j h_j^2} e^{-\theta_{j'} h_{j'}^2}$$

- Composite, unequal weights

$$R^{\mathrm{C-UW}}(\mathbf{x}, \mathbf{x}') = \lambda_1 \cdot \frac{1}{d} \sum_{j=1}^{d} e^{-\theta_j h_j^2} + \lambda_2 \cdot \frac{2}{d(d-1)} \sum_{j=1}^{d} \sum_{j'=j+1}^{d} e^{-\theta_j h_j^2} e^{-\theta_{j'} h_{j'}^2}$$

(λ_1 and λ_2 estimated subject to $\lambda_1 + \lambda_2 = 1$)

Franke's Function: Composite Correlation Structures (Additive and All 2-Input Interactions)



Normalized RMSE versus n and log-log plot to show rate of convergence

Composite Select Correlation Structures (Additive and Some 2-Input Interactions)

- Composite with select 2-input interactions, equal weights

$$R^{\mathrm{CS-EW}}(\mathbf{x}, \mathbf{x}') = \frac{1}{2} \cdot \frac{1}{d} \sum_{j=1}^{d} e^{-\theta_j h_j^2} + \frac{1}{2} \cdot \frac{1}{|B|} \sum_{\{j,j'\} \in B} e^{-\theta_j h_j^2} e^{-\theta_{j'} h_{j'}^2}$$

- Composite with select 2-input interactions, unequal weights

$$R^{\mathrm{CS-UW}}(\mathbf{x}, \mathbf{x}') = \lambda_1 \cdot \frac{1}{d} \sum_{j=1}^{d} e^{-\theta_j h_j^2} + \lambda_2 \cdot \frac{1}{|B|} \sum_{\{j,j'\} \in B} e^{-\theta_j h_j^2} e^{-\theta_{j'} h_{j'}^2}$$

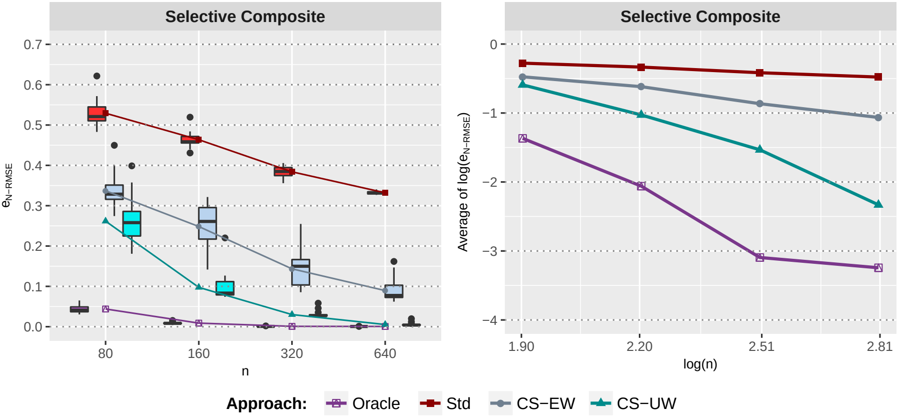
Choosing the 2-Input Interactions: FANOVA (Schonlau and Welch, 2006)

- Choose the interaction effects for the 8-dimensional Franke's function data adaptively
- Functional ANOVA (FANOVA) based on a training set with $n = 640$
- Percent contributions to total variance of the function

Effect	%	Effect	%	Effect	%
x_1	8.1	x_2	15.8	$x_1 \times x_2$	2.4
x_3	7.5	x_4	13.5	$x_3 \times x_4$	2.3
x_5	7.9	x_6	15.1	$x_5 \times x_6$	2.4
x_7	7.4	x_8	14.5	$x_7 \times x_8$	2.3

- These effects account for 99.2% of the variation (others negligible)
- $B = \{\{1,2\},\{3,4\},\{5,6\},\{7,8\}\}$.

Franke's Function: Select Composite (Main Effects and Select 2-Input Interaction Effects)



Normalized RMSE versus *n* and log-log plot to show rate of convergence

Bad Boys: The Math

- High dimensionality: $d \gg 1$ inputs are active, say $x_1, \ldots, x_d$
- Nonlinearity: their θ_j are large
- What does the product correlation function do?

$$
\begin{aligned}
R^{\mathrm{Std}}(\mathbf{x}, \mathbf{x}') = \prod_{j=1}^{d} R(h_j) = \prod_{j=1}^{d} e^{-\theta_j h_j^2} &= \prod_{j=1}^{d} (1 - \theta_j h_j^2 + \frac{1}{2}(\theta_j h_j^2)^2 + \cdots) \\
&= 1 - \theta_1 h_1^2 + \frac{1}{2}(\theta_1 h_1^2)^2 + \text{similar terms for } x_2, \text{ etc.} \\
&+ \theta_1 h_1^2 \theta_2 h_2^2 + \text{similar terms for } x_3, x_4, \text{ etc.} \\
&+ \cdots \pm \prod_{j=1}^{d} \theta_j h_j^2
\end{aligned}
$$

- The higher-order products will not disappear if the $\theta_j \gg 0$ unless all $h_j \simeq 0$ (huge sample size)

Conclusions

- With high dimensionality and nonlinearity, the product correlation function automatically introduces high-order interactions, needed or not

Conclusions

- With high dimensionality and nonlinearity, the product correlation function automatically introduces high-order interactions, needed or not
- Proof of concept
- Effect sparsity, hierarchy and heredity applied to computer experiments

Conclusions

- With high dimensionality and nonlinearity, the product correlation function automatically introduces high-order interactions, needed or not
- Proof of concept
- Effect sparsity, hierarchy and heredity applied to computer experiments
- Michalewicz test function: highly nonlinear, moderate dimension
  - Useful accuracy with a main-effects model
- New Franke function: nonlinear, moderate dimension
  - More accuracy with limited, 2-input interactions

Conclusions

- With high dimensionality and nonlinearity, the product correlation function automatically introduces high-order interactions, needed or not
- Proof of concept
- Effect sparsity, hierarchy and heredity applied to computer experiments
- Michalewicz test function: highly nonlinear, moderate dimension
  - Useful accuracy with a main-effects model
- New Franke function: nonlinear, moderate dimension
  - More accuracy with limited, 2-input interactions
- Work by Alexi Rodríguez to extend these ideas to real applications is also taking account of the science (Dennis Lin's paper in JQT)

THANK YOU! ⌣̈

References

Ba, S. and Joseph, V. R. (2012). Composite Gaussian process models for emulating expensive functions. *The Annals of Applied Statistics*, 6(4):1838–1860.

Box, G. and Meyer, R. D. (1985). Some new ideas in the analysis of screening designs. *Journal of Research of the National Bureau of Standards*, 90(6):495–500.

Chen, H., Loeppky, J. L., Sacks, J., and Welch, W. J. (2016). Analysis methods for computer experiments: How to assess and what counts? *Statistical Science*, 31(1):40–60.

Dancik, G. M. (2013). *mlegp: Maximum Likelihood Estimates of Gaussian Processes*. R package version 3.1.5.

Franke, R. (1979). *A Critical Comparison of Some Methods for Interpolation of Scattered Data*. Final report. Defense Technical Information Center.

Haaland, B. and Qian, P. Z. G. (2011). Accurate emulators for large-scale computer experiments. *The Annals of Statistics*, 39(6):2974–3002.

Hamada, M. and Wu, C. F. J. (1992). Analysis of designed experiments with complex aliasing. *Journal of Quality Technology*, 24(3):130–137.

Schonlau, M. and Welch, W. J. (2006). Screening the input variables to a computer model via analysis of variance and visualization. In Dean, A. and Lewis, S., editors, *Screening: Methods for Experimentation in Industry, Drug Discovery, and Genetics*, pages 308–327. Springer, New York.

Wu, C. F. J. and Hamada, M. S. (2009). *Experiments: Planning, Analysis, and Optimization*. Wiley, New Jersey, second edition.

Yates, F. (1937). *The Design and Analysis of Factorial Experiments*. Technical Communication No. 35, Imperial Bureau of Soil Science, Farnham Royal.