

Robust simulation design for generalised linear models in conditions of heteroscedasticity or correlation

Andrew Gill^{*1}, David J. Warne^{2,3}, Antony M. Overstall⁴, Clare McGrory², and James M. McGree^{2,3}

¹Joint and Operations Analysis Division, Defence Science and Technology Group, Edinburgh, Australia

²School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia

³Centre for Data Science, Queensland University of Technology, Brisbane, Australia

⁴Mathematical Sciences, University of Southampton, Southampton, UK

December 21, 2022

Abstract

A meta-model of the input-output data of a computationally expensive simulation is often employed for prediction, optimization, or sensitivity analysis purposes. Fitting is enabled by a designed experiment, and for computationally expensive simulations, the design efficiency is of importance. Heteroscedasticity in simulation output is common, and it is potentially beneficial to induce dependence through the reuse of pseudo-random number streams to reduce the variance of the meta-model parameter estimators. In this paper, we develop a computational approach to robust design for computer experiments without the need to assume independence or identical distribution of errors. Through explicit inclusion of the variance or correlation structures into the meta-model distribution, either maximum likelihood estimation or generalized estimating equations can be employed to obtain an appropriate Fisher information matrix. Robust designs can then be computationally sought which maximize some relevant summary measure of this matrix, averaged across a prior distribution of any unknown parameters.

1 Introduction

The fitting of a model to a sample of input-output data of a computationally expensive simulation is an important task in simulation analytics (Santner et al., 2003). This model of a model (meta-model) can then be efficiently employed for prediction, optimization, or sensitivity analysis purposes. Sensitivity analyses are often well served by low-order polynomial (usually quadratic) meta-models, as they enable the characterisation of impact (main effects), synergies (two factor interactions) and diminishing returns (squared terms) (Gill et al., 2018; Sanchez et al., 2012).

Fitting of the meta-model is enabled by a designed experiment, and for computationally expensive (often stochastic) simulations, the efficiency of the design is of importance (Fedorov, 1972). For

^{*}To whom correspondence should be addressed. E-mail: andrew.gill@defence.gov.au

linear meta-models, factorial-based designs (often fractional and supplemented with central and axial points if fully quadratic) are typically prescribed (Montgomery, 2012), as these are efficient under D -optimality if the typically assumed condition of independent and identically distributed (iid) errors holds.

Kleijnen (2015) is perhaps the seminal text on experimental design for simulation, and discusses the implications of departures from iid conditions for the analysis of linear meta-models, which Gill (2019) illustrates. However, while the assumption of independence can actually be assured in simulation by employing unique pseudo random number (PRN) streams at each design point, this overlooks an important variance reduction (design efficiency) opportunity. Schruben and Margolin (1978) were the first to devise a design efficient PRN assignment strategy for linear meta-models and (generally) factorial-based designs (Gill (2021) illustrates with a simple example).

However, Kleijnen (2015) is relatively silent on the question of design when iid conditions do not hold for linear meta-models (“*the literature pays little attention to the derivation of alternative designs for cases with heterogeneous output variances*” and “*the literature pays no attention to the derivation of alternative designs for situations with common random numbers (CRN)*”). Furthermore, simulation outputs are often discrete and sometimes only binary, so the broader range of generalized linear (meta-)models (GLMs) are typically required (i.e., linear, Poisson, and logistic) (Dunn and Smyth, 2018). Woods et al. (2006) point out that the design efficiency for GLMs depends on the regression parameters yet to be estimated, so that robust designs are often sought by computational optimization.

In this paper, we seek to bring to the attention of the simulation analytics community literature which address some of these design-related gaps. In particular, we illustrate in some detail design construction for linear meta-models in the presence of heteroscedasticity (drawing on Atkinson and Cook (1995)) and GLMs in the presence of correlation (Woods and van de Ven, 2011), before concluding with a proof of concept for the idea of jointly optimizing both the design and PRN assignment for linear meta-models.

2 Robust design construction for GLM

2.1 GLM designs

In the GLM framework, for each input of q factors $\mathbf{x}_i = [x_{i,1}, x_{i,2}, \dots, x_{i,q}] \in \mathbb{R}^q \quad i = 1, \dots, n$ we have a simulation response Y_i with a probability mass/density function $p(y_i)$ assumed to come from the exponential family of distributions and where there is an appropriate link function $g(\cdot)$ such that $g(\mathbb{E}_{\mathbf{Y}}[Y_i|\mathbf{x}_i]) = \mathbf{f}^T(\mathbf{x}_i)\boldsymbol{\beta}$. Here $\boldsymbol{\beta} = [\beta_0, \beta_1, \dots, \beta_{d-1}]^T$ is a column vector of d ($q < d \leq n$) unknown parameters and $\mathbf{f}^T(\mathbf{x}_i) : \mathbb{R}^q \rightarrow \mathbb{R}^d$ is a row vector of d terms that may include first order and higher order interactions of the q input factors.

The goal is to choose the set of n design points $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times q}$ to efficiently estimate $\boldsymbol{\beta}$. Minimizing the approximate volume of the covariance ellipsoid of the maximum likelihood estimator of $\boldsymbol{\beta}$ is equivalent to maximizing the determinant of the Fisher information matrix (hence called D -optimal)

$$\mathbf{X}^* = \underset{\mathbf{X}}{\operatorname{argmax}} |I_{\mathbf{X}}(\boldsymbol{\beta})|, \quad I_{\mathbf{X}}(\boldsymbol{\beta})_{j,k} = -\mathbb{E}_{\mathbf{Y}} \left[\frac{\partial^2 \ell(\mathbf{Y}, \mathbf{F}, \boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} \middle| \boldsymbol{\beta} \right] \quad (1)$$

where $\mathbf{F} = [\mathbf{f}^T(\mathbf{x}_1), \mathbf{f}^T(\mathbf{x}_2), \dots, \mathbf{f}^T(\mathbf{x}_n)]^T$ is an $n \times d$ matrix and $\ell(\mathbf{y}, \mathbf{F}, \boldsymbol{\beta}) = \sum_{i=1}^n \log p(y_i | \mathbf{f}^T(\mathbf{x}_i), \boldsymbol{\beta})$ is the (assumed twice differentiable) log-likelihood for the observations $\mathbf{y} = [y_1, y_2, \dots, y_n]^T$ at the design points \mathbf{X} given the parameters $\boldsymbol{\beta}$.

Using second order partial derivatives of $\log p(y_i | \mathbf{f}^T(\mathbf{x}_i), \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, then taking expectations with respect to Y_i , we obtain the following expression for the expected Fisher information

matrix

$$I_{\mathbf{X}}(\boldsymbol{\beta}) = \mathbf{F}^T \mathbf{P} \mathbf{F} \quad (2)$$

where $\mathbf{P} = \text{diag} \left(1 / \left[g' \left(\mathbb{E}_{\mathbf{Y}} [Y_i | \mathbf{x}_i] \right)^2 \text{Var}_{\mathbf{Y}} [Y_i | \mathbf{x}_i] \right] \right)$ which is a known function of \mathbf{F} and $\boldsymbol{\beta}$ for the relevant exponential family distribution using in the GLM.

Designs based on (1) and (2) assume a fixed number of design points n (an exact design). If instead we ascribe to \mathbf{x}_i a weight $0 \leq w_i \leq 1$ (with $\sum_{i=1} w_i = 1$ thus representing how sampling effort is distributed across design points) and relabel $\mathbf{x}_i = [x_{i,1}, x_{i,2}, \dots, x_{i,k}, w_i]$, so that $\mathbf{X} \in \mathbb{R}^{n \times q+1}$, then the approximate design problem is to find $\mathbf{X}^* = \text{argmax}_{\mathbf{X}} |I_{\mathbf{X}}(\boldsymbol{\beta})|$ where $I_{\mathbf{X}}(\boldsymbol{\beta}) = \mathbf{F}^T \mathbf{W} \mathbf{P} \mathbf{F}$ where $W_{ii} = w_i$. From this approximate design, an exact design of a particular size can be generated by sampling according to the weights w_i^* .

2.2 Robust design

Obviously the requirement to know $\boldsymbol{\beta}$ *a priori* is not useful for finding designs for estimating $\boldsymbol{\beta}$. A common approach to remove the $\boldsymbol{\beta}$ dependency is to average (some monotonic function of) the optimality criterion across a prior distribution $\pi(\boldsymbol{\beta})$ of possible values of $\boldsymbol{\beta}$

$$\mathbf{X}^* = \text{argmax}_{\mathbf{X}} \int \log(|I_{\mathbf{X}}(\boldsymbol{\beta})|) \pi(\boldsymbol{\beta}) d\boldsymbol{\beta} \quad (3)$$

where the logarithm of the determinant is often used for numerical stability purposes. We call this pseudo-Bayesian approach (Chaloner and Verdinelli, 1995; Englezou, 2018) robust design, as it is robust to misspecification of the parameters (though not the meta-model - see Section 6). The prior can be based on previous investigations or subject matter expertise, or a non-informative probability distribution if required.

Often, the integral in (3) is not analytically tractable, so numerical integration is required. Quadrature rules are possible but are more cumbersome in higher dimensions, so here we use a direct Monte Carlo estimator, so that

$$\mathbf{X}^* \approx \text{argmax}_{\mathbf{X}} \frac{1}{M} \sum_{m=1}^M \log(|I_{\mathbf{X}}(\boldsymbol{\beta}_m)|),$$

where $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_M$ are iid draws from the prior $\pi(\boldsymbol{\beta})$.

Robust design using a Monte Carlo estimate of the expected Fisher information requires the maximization of a random quantity with variance of $\mathcal{O}(1/M)$. Many standard non-linear optimization algorithms, such as Levenberg-Marquardt (Levenberg, 1944; Marquardt, 1963), cannot handle random variables in the function to be optimized. Instead, we apply simulated annealing, which is a probabilistic optimization technique (Kirkpatrick et al., 1983). Other methods for stochastic optimization such as the Approximate Coordinate Exchange algorithm (Overstall and Woods, 2017) can be more efficient for more complex functions, but simulated annealing is sufficient here.

3 Robust designs for departures from iid conditions

3.1 Linear meta-model in the presence of heteroscedasticity

Consider a $q = 2$ design problem $\mathbf{x} = [x_1, x_2] \in [-1, 1]^2$ for the full second-order polynomial linear meta-model, so $\mathbf{f}^T(\mathbf{x}_i) = [1, x_{i,1}, x_{i,2}, x_{i,1}x_{i,2}, x_{i,1}^2, x_{i,2}^2]$ and $g(\cdot)$ is the identity function, but where $Y_i \sim N(\mathbf{f}^T(\mathbf{x}_i)\boldsymbol{\beta}, \sigma^2 v(\mathbf{x}_i))$ with $v(\mathbf{x}_i) = \exp(\mathbf{x}_i \boldsymbol{\alpha} [1 + 2\mathbf{x}_i \boldsymbol{\alpha}])$. Here $\boldsymbol{\alpha}$ is a column vector with the same dimension as \mathbf{x}_i but otherwise unknown. Thus, the Y_i are independent, but $\|\boldsymbol{\alpha}\|_2^2$ controls the degree of heteroscedasticity.

To find robust designs, we need the expected information matrix for this meta-model. Since Y_i is normally distributed, it's log-likelihood at design point \mathbf{x}_i is

$$\ell_i(y_i, \mathbf{f}^\top(\mathbf{x}_i), \boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{-(y_i - \mathbf{f}^\top(\mathbf{x}_i)\boldsymbol{\beta})^2}{2\sigma^2 \exp(\mathbf{x}_i\boldsymbol{\alpha}[1 + 2\mathbf{x}_i\boldsymbol{\alpha}])} - \frac{\mathbf{x}_i\boldsymbol{\alpha}[1 + 2\mathbf{x}_i\boldsymbol{\alpha}]}{2} - \log(\sqrt{2\pi}\sigma)$$

and it is relatively easy to show that $\mathbb{E}_{Y_i} \left[\frac{\partial^2 \ell_i}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\alpha}_k} \right] = 0$ given $\mathbb{E}_{\mathbf{Y}} [Y_i | \mathbf{x}_i] = \mathbf{f}^\top(\mathbf{x}_i)\boldsymbol{\beta}$, which means the Fisher information matrix in (1) will be block diagonal with two blocks; one for $\boldsymbol{\beta}$ and the other for $\boldsymbol{\alpha}$. For these

$$\begin{aligned} \frac{\partial^2 \ell_i}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_k} &= \frac{-\mathbf{X}_{i,j}\mathbf{X}_{i,k}}{\sigma^2 \exp(\mathbf{x}_i\boldsymbol{\alpha}[1 + 2\mathbf{x}_i\boldsymbol{\alpha}])} \\ \frac{\partial^2 \ell_i}{\partial \boldsymbol{\alpha}_j \partial \boldsymbol{\alpha}_k} &= -\frac{1}{2}\mathbf{X}_{i,j}\mathbf{X}_{i,k} \left(4 - [4 - (1 + 4\mathbf{x}_i\boldsymbol{\alpha})^2] \frac{(y_i - \mathbf{f}^\top(\mathbf{x}_i)\boldsymbol{\beta})^2}{\sigma^2 \exp(\mathbf{x}_i\boldsymbol{\alpha}[1 + 2\mathbf{x}_i\boldsymbol{\alpha}])} \right). \end{aligned}$$

Clearly, $I_{\mathbf{X}}(\boldsymbol{\beta})$ takes the form (2) with $\mathbf{P}_{ii} = (\sigma^2 \exp(\mathbf{x}_i\boldsymbol{\alpha}[1 + 2\mathbf{x}_i\boldsymbol{\alpha}]))^{-1} = (\sigma^2 v(\mathbf{x}_i))^{-1}$ and given $\mathbb{E}_{\mathbf{Y}} [(y_i - \mathbf{f}^\top(\mathbf{x}_i)\boldsymbol{\beta})^2 | \mathbf{x}_i] = \sigma^2 \exp(\mathbf{x}_i\boldsymbol{\alpha}[1 + 2\mathbf{x}_i\boldsymbol{\alpha}])$ we see that $I_{\mathbf{X}}(\boldsymbol{\alpha}) = \mathbf{X}^\top \mathbf{Q} \mathbf{X}$ with $\mathbf{Q}_{ii} = \frac{1}{2}[1 + 4\mathbf{x}_i\boldsymbol{\alpha}]^2$. We note that for linear meta-models, \mathbf{P} and \mathbf{Q} do not depend on $\boldsymbol{\beta}$. This accords with Atkinson and Cook (1995) and their original derivation which showed that the information expected to be obtained about $\boldsymbol{\beta}$ based on the i -th design point is given by $\mathbf{f}^\top(\mathbf{x}_i)\mathbf{f}(\mathbf{x}_i)/(\sigma^2 v(\mathbf{x}_i))$ while for $\boldsymbol{\alpha}$ it is presented as $\mathbf{J}^\top \mathbf{J}$, where $\mathbf{J} = (1 + 4\mathbf{x}_i\boldsymbol{\alpha})\mathbf{x}_i/\sqrt{2}$.

As a means of comparison, the prior considered in Atkinson and Cook (1995) for $\boldsymbol{\alpha}$ placed equal mass on the following five values: $[1, 0]$, $[0.75, 0.25]$, $[0.5, 0.5]$, $[0.25, 0.75]$ and $[0, 1]$. The motivation is that these values span the directions in which the variance increases with x_1 and x_2 , and that there is no prior knowledge to suggest which direction is more likely than another. Figure 1 shows the local D -optimal designs for each unique value of $\boldsymbol{\alpha}$ along with the robust design, assuming the mean is known (thus focusing on $I_{\mathbf{X}}(\boldsymbol{\alpha})$).

Simulated annealing was employed to locate each design including the design weights. To do so, the optimisation was initialised with a random selection of design points and design weights with a relatively large value of n . Throughout the optimisation, if some weights approached zero, then the corresponding design points were removed, which is why some optimal designs have different numbers of unique experimental runs.

Notably, these designs are very similar to those presented in Figure 5 of Atkinson and Cook (1995) (including the weights w_i , not shown here). For the locally optimal designs, symmetry about $\boldsymbol{\alpha}$ is observed. This is expected given how $\boldsymbol{\alpha}$ and \mathbf{x} exist in the model. The robust design resembles a compromise between the designs found for each value of $\boldsymbol{\alpha}$ with the largest experimental effort being assigned to $\mathbf{x} = [1, 1]$. Further, the points for $x_1 = 1$ and $x_2 = 1$ align with design points selected for different values of $\boldsymbol{\alpha}$. Lastly, there is an inner point placed near $\mathbf{x} = [0, 0]$ which appears to be a compromise between the additional design point found at extreme values for $\boldsymbol{\alpha}$.

3.2 Logistic meta-model in the presence of correlation

3.2.1 Fisher information matrix via generalized estimating equations

Now consider a Bernoulli response Y_i , so that $P(Y_i = 1) = p_i = \mathbb{E}_{\mathbf{Y}} [Y_i | \mathbf{x}_i]$ and $g(\cdot)$ is the logit function, with $q = 3$ input factors and their pairwise interactions

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,1}x_{i,2} + \beta_5 x_{i,1}x_{i,3} + \beta_6 x_{i,2}x_{i,3} + \boldsymbol{\varepsilon}_i \quad (4)$$

but where we have added latent random variables $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R})$ with a general $n \times n$ covariance matrix $\mathbf{R}_{i,j} = R(\mathbf{x}_i, \mathbf{x}_j)$. Unlike the linear meta-model, the $\text{logit}(\cdot)$ introduces non-linear terms into the

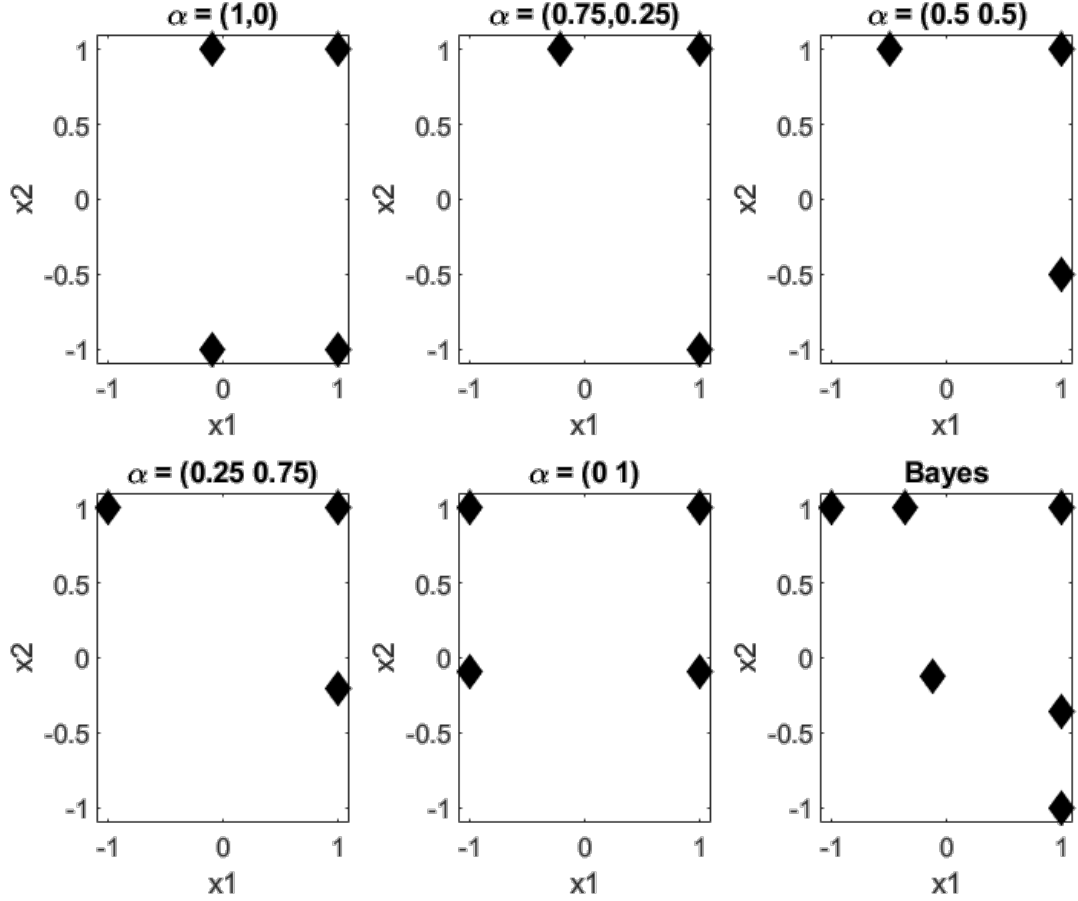


Figure 1: D -optimal designs for various values of α and the robust (Bayes) design.

expression for the log-likelihood, which will render the analytical integration over $\boldsymbol{\varepsilon}$ to obtain the marginal likelihood impossible. Therefore, it is not possible to obtain an exact analytic expression to the expected Fisher information matrix for the model given in (4).

However, following Woods and van de Ven (2011), we can obtain an approximation using generalized estimating equations (GEE) (see Liang and Zeger (1986) for details). For the logistic GLM with correlations, the GEE leads to the following approximation (in the weighted design context)

$$I_{\mathbf{X},\mathbf{R}}(\boldsymbol{\beta}) \approx \mathbf{F}^T(\mathbf{W}\mathbf{P})^{1/2}\mathbf{R}^{-1}(\mathbf{W}\mathbf{P})^{1/2}\mathbf{F}, \quad (5)$$

where the dependence on $\boldsymbol{\beta}$ is observed through \mathbf{P} with $P_{ii} = p_i(1-p_i) = \exp(\mathbf{f}^T(\mathbf{x}_i)\boldsymbol{\beta}) / (1 + \exp(\mathbf{f}^T(\mathbf{x}_i)\boldsymbol{\beta}))^{-2}$, and $\mathbf{W}_{ii} = w_i$ with weight $0 \leq w_i \leq 1$ (with $\sum_{i=1} w_i = 1$ thus representing how sampling effort is distributed across design points) and relabel $\mathbf{x}_i = [x_{i,1}, x_{i,2}, \dots, x_{i,k}, w_i]$, so that $\mathbf{X} \in \mathbb{R}^{n \times q+1}$.

For the purposes of this study, we assume constant (homoscedastic) variance $R(\mathbf{x}_i, \mathbf{x}_i) = \sigma^2$. The covariance structures we consider are as follows (for $i \neq j$).

- *Independent*: The standard assumption in which (5) reduces to (2), i.e. $R(\mathbf{x}_i, \mathbf{x}_j) = 0$.
- *Constant*: All observations are equally correlated with each other, i.e. $R(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \rho$.
- *Auto-regressive*: The observation index is treated as a time index, i.e. $R(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \rho^{|i-j|}$.
- *Distance-kernel*: Isotropic spatial correlation between design points, i.e. $R(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \rho e^{-\frac{1}{4} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2}$.

Here $\rho \in [0, 1]$ is a correlation parameter (for now considering only positive correlations). Each correlation structure could, in principle, be valid for a specific computer simulation experiment. If this structure is known *a priori*, then that structure should be used for the design. However, for most experiments, the correlation structure is not known. Therefore, we seek to understand the efficiency of each correlation assumption under misspecification.

3.2.2 Evaluating design efficiency under misspecification

To consider the question of design efficiency for the logistic GLM with correlations (4), we perform a simulation study. For each of the above correlation structures we obtain a robust design using our computational approach, and assess the efficiency under misspecification of that correlation. We define some notation to express this comparison more formally. Let $\mathbf{X}^*(\mathbf{R})$ denote a robust design under the D -optimality criterion (3) using the GEE approximation (5) with covariance matrix \mathbf{R} . Then define

$$\mathcal{J}(\mathbf{X}, \mathbf{R}) = \int |I_{\mathbf{X}, \mathbf{R}}(\boldsymbol{\beta})|^{1/d} \pi(\boldsymbol{\beta}) d\boldsymbol{\beta} \approx \frac{1}{M} \sum_{m=1}^M |I_{\mathbf{X}, \mathbf{R}}(\boldsymbol{\beta}_m)|^{1/d}$$

where the d -th root is routinely used to allow fair comparisons between designs. The ratio $\mathcal{J}[\mathbf{X}_1, \mathbf{R}] / \mathcal{J}[\mathbf{X}_2, \mathbf{R}]$ gives the D -efficiency of a design \mathbf{X}_1 relative to a reference design \mathbf{X}_2 given a covariance matrix \mathbf{R} (Woods and van de Ven, 2011). The D -efficiency can be interpreted as the amount of additional experimental effort needed, whereby if D -efficiency is 0.5, then you would need to run the design twice to obtain as much information as the optimal design.

Now consider two covariance matrices \mathbf{R}_1 and \mathbf{R}_2 , then $\mathbf{X}^*(\mathbf{R}_1)$ denotes the robust design assuming \mathbf{R}_1 , and similarly $\mathbf{X}^*(\mathbf{R}_2)$ is robust assuming \mathbf{R}_2 . It follows, that

$$\text{Misspecification } D\text{-Efficiency}(\mathbf{R}_1, \mathbf{R}_2) = \frac{\mathcal{J}[\mathbf{X}^*(\mathbf{R}_1), \mathbf{R}_2]}{\mathcal{J}[\mathbf{X}^*(\mathbf{R}_2), \mathbf{R}_2]}, \quad (6)$$

represents the D -efficiency of a design using the misspecified \mathbf{R}_1 when \mathbf{R}_2 was the true covariance.

We evaluate this misspecification efficiency (6) for each pair of covariance functions and do this for a range of $\rho \in [0.05, 0.95]$ to investigate how the efficiency depends on the correlation strength. For each design simulation, we optimize n weighted design points for the $q = 3$ factor model using the robust design expected utility estimated with $M = 1,000$ prior samples. When evaluating the final efficiency losses we use a more precise Monte Carlo estimate with $M = 20,000$. The resulting efficiency as a function of correlation strength is provided for each pair of covariance structures in Figure 2.

Note that the efficiencies > 1 suggest too small a value of the Monte Carlo sampling rate ($M = 1,000$). However, to produce Figure 2 required 25 robust designs (8 values of ρ for each of the 3 covariance options dependent on ρ with an additional independent case), with each design run costing approximately 16 hours of CPU time on a Intel Xeon Gold 6140 processor (total of approximately 400 CPU hours distributed over 18 cores) which motivated the chosen value of M for this study.

Several important patterns are observed in Figure 2. Firstly, there is almost no penalty for assuming a correlation structure when independence is valid. That is, the efficiency of constant, auto-regressive or distance correlation relative to independence is > 0.9 (Figure 2, blue lines). A similar insensitivity is apparent for misspecification relative to constant correlation. However, the situation becomes quite different when considering misspecification relative to auto-regressive or distance-based correlation. In both cases, the penalty of misspecification increases as the correlation strength, ρ , increases. Assuming distance correlation when auto-regressive is true performs better overall than the converse relationship. However, when $\rho > 0.7$, constant correlation starts to be the better assumption under misspecification by auto-regressive or distance correlation.

Thus, if sufficient knowledge is available to prescribe a correlation assumption with certainty, then this will always be the best choice. Beyond this unrealistic case, one clear result is that the independent assumption should only be used if the risks of misspecification is low. The same can mostly be

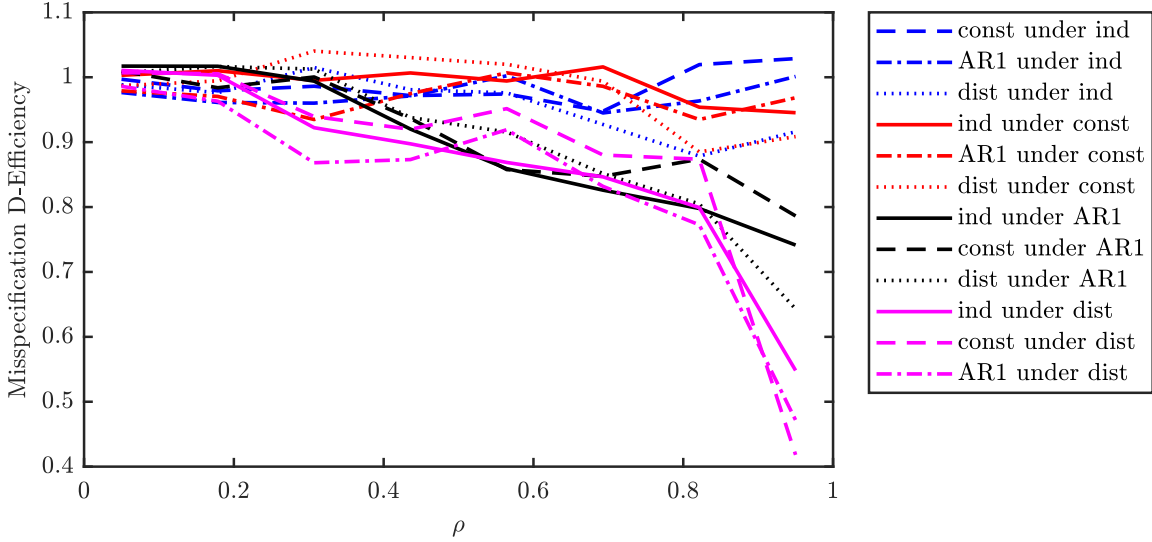


Figure 2: The efficiency of designs under different combinations of assumed (\mathbf{R}_1) under true (\mathbf{R}_2) correlation assumptions plotted against correlation strength.

said for the constant assumption, unless the correlation strength is higher, in which case the more specific structural distinctions between auto-regressive and distance correlation become apparent. The choice of auto-regressive or distance based correlation assumptions do not reduce the quality of the design substantially if independent or constant correlations would have been also been valid choices. However, the choice of auto-regressive and distance correlation is more complex and depends on the correlation strength. If one can rule out auto-regressive correlation, that represents temporal correlation, then this causes few problems and distance correlation should be used. However, if it is unclear if distance or auto-regressive are possibilities, then additional exploration is needed. In general we arrive at the following recommendations.

1. If $R(\mathbf{x}_i, \mathbf{x}_j)$ is known, use this in the design process.
2. If $R(\mathbf{x}_i, \mathbf{x}_j)$ is uncertain, but auto-regressive correlation can be excluded, then use distance-kernel correlation.
3. If $R(\mathbf{x}_i, \mathbf{x}_j)$ is completely uncertain, some understanding of the range of ρ is required. If $\rho \leq 0.7$ then distance-kernel correlation is more robust, otherwise constant correlation is more robust.

We also investigated the qualitative differences in the design patterns for the various correlation structures and values of correlation strength ρ . The example spatial patterns shown in Figure 3 correspond to a view along the x_1 -axes (the other axes views are very similar qualitatively).

Given the efficiency results, it is not surprising that the design patterns look similar for different values of ρ . However, for a fixed ρ we can observe some differences between designs under different correlation assumptions. Both the independent and constant correlation cases are characterised with fewer points, but with relatively constant weights, however, auto-regressive and distance correlation tend to have more points with lower weights.

4 Joint optimization of design and PRN assignment

Sections 2 and 3 describe design construction for GLMs where iid conditions may not be present. Unlike physical experiments, where dependence or correlation may arise due to unavoidable constraints

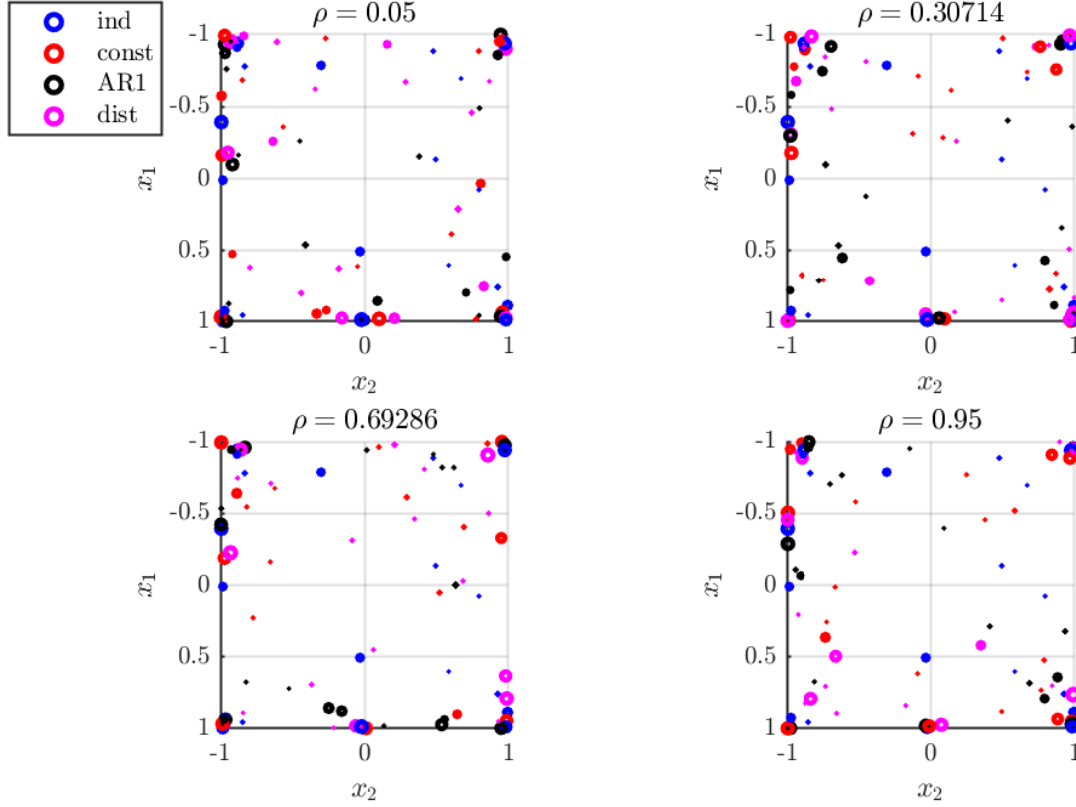


Figure 3: The weighted design points under different correlation assumptions. The weight of a design point is represented by the circle radius. View is along the x_1 -axis.

and nuisance blocking effects need to be accounted for, simulation experiments can guarantee independence by using different PRN streams for each design point. However, simulation experiments can conversely induce correlations by the use of CRN. Schruben and Margolin (1978) were among the first to clearly illustrate how doing so can improve the D-efficiency of a given design, and provided an assignment strategy for mostly factorial-based designs. Of interest in this section is the merging of that idea with the design construction approaches of the previous sections.

4.1 Linear meta-model in the presence of correlation

Suppose there are $2g$ PRN streams given by g streams (denoted R_1, \dots, R_g) and their antitheses (denoted $\bar{R}_1, \dots, \bar{R}_g$). Denote the $2g$ streams R_1, \dots, R_{2g} with $R_{g+j} = \bar{R}_j$, for $j = 1, \dots, g$ and let $b(R_j)$ be the random block effect associated with PRN stream R_j . Suppose now that design point \mathbf{x}_i is assigned PRN stream $R_{k(i)}$ for some assignment strategy $k(\cdot)$. Let \mathbf{Z} be the $n \times 2g$ matrix with

$$\mathbf{Z}_{ih} = \begin{cases} 1 & \text{if } R_h \text{ is used for experimental run } i; \\ 0 & \text{otherwise,} \end{cases}$$

and $\mathbf{b} = [b(R_1), \dots, b(R_{2g})]^T$ be the $2g \times 1$ column vector of unique random block effects. Then, $\boldsymbol{\gamma} = \mathbf{Z}\mathbf{b} = [b(R_{k(1)}), \dots, b(R_{k(n)})]^T$ is the column vector of n random block effects as assigned in the experiment. Now $E(\boldsymbol{\gamma}) = \mathbf{0}_n$ and $\text{Var}(\boldsymbol{\gamma}) = \mathbf{R}$, where \mathbf{R} is the $n \times n$ matrix with

$$R_{i,j} = \text{Cov}(b(R_{k(i)}), b(R_{k(j)})) = \begin{cases} \sigma^2 \rho_+ & \text{if } k(i) = k(j); \\ -\sigma^2 \rho_- & \text{if } |k(i) - k(j)| = g; \\ 0 & \text{otherwise,} \end{cases}$$

where $\rho_- > 0$ and $\rho_+ > 0$ are unknown.

Here ρ_+ and $-\rho_-$ are positive and negative correlations induced by using the same PRN stream or its antithesis for experimental points i and j . Following Schruben and Margolin (1978), the model is

$$Y_i = \mathbf{f}^T(\mathbf{x}_i)\boldsymbol{\beta} + \gamma_i + \varepsilon_i \quad (7)$$

where $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2(1 - \rho_+)\mathbf{I}$. For a linear meta-model, we can use the Ordinary Least Squares estimator $\hat{\boldsymbol{\beta}} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{y}$, which can be shown to have variance (under (7))

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{V}\mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1} \text{ where } \mathbf{V} = (1 - \rho_+)\mathbf{I} + \mathbf{Z}\mathbf{R}\mathbf{Z}^T.$$

4.2 Optimal blocked designs

Optimal design for blocked experiments has been considered previously (see, for example, Chapters 7 and 8 of Goos and Jones (2011) or Chapter 15 of Donev et al. (2007)). What is different here is that there is positive correlation between elements of \mathbf{b} whereas these are usually assumed to be independent. Design specification here involves both the choice of $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$ and the PRN allocation $\mathbf{k} = [k(1), \dots, k(n)]$ where $k(i) \in \{1, \dots, 2g\}$.

While before we used the (determinant of the) Fisher information matrix, as the goal is to minimize the volume of the covariance ellipsoid, here we can directly and equivalently use the (log of, for numerical stability) determinant of $\text{Var}(\hat{\boldsymbol{\beta}})$

$$(\mathbf{X}^*, \mathbf{k}^*) = \underset{(\mathbf{X}, \mathbf{k})}{\text{argmin}} (\log |\mathbf{F}^T\mathbf{V}\mathbf{F}| - 2\log |\mathbf{F}^T\mathbf{F}|).$$

However, \mathbf{V} depends on unknowns ρ_- and ρ_+ through \mathbf{R} . As before, we instead seek a robust design under a joint prior distribution $\pi(\rho_-, \rho_+)$ with domain $[0, 1]^2$

$$(\mathbf{X}^*, \mathbf{k}^*) = \underset{(\mathbf{X}, \mathbf{k})}{\text{argmin}} \left(\int_0^1 \int_0^1 \log |\mathbf{F}^T\mathbf{V}\mathbf{F}| \pi(\rho_+, \rho_-) d\rho_- d\rho_+ - 2\log |\mathbf{F}^T\mathbf{F}| \right) \quad (8)$$

and as the integral in (8) is not typically available in closed form, it is evaluated here using a 2-dimensional Gauss-Legendre quadrature rule (see, for example, Weiser (2016))

$$(\mathbf{X}^*, \mathbf{k}^*) \approx \underset{(\mathbf{X}, \mathbf{k})}{\text{argmin}} \left(\sum_{m=1}^M \omega_m \log |\mathbf{F}^T\mathbf{V}_m\mathbf{F}| - 2\log |\mathbf{F}^T\mathbf{F}| \right)$$

where $\omega_1, \dots, \omega_M$ are the quadrature weights and \mathbf{V}_m is \mathbf{V} evaluated at the corresponding quadrature nodes $(\rho_-^{(m)}, \rho_+^{(m)})$. For illustrative purposes it suffices here to use a rudimentary joint optimization

$$(\mathbf{X}^*, \mathbf{k}^*) \approx \underset{\mathbf{X}}{\text{argmin}} \left(\min_{\mathbf{k}} \left(\sum_{m=1}^M \omega_m \log |\mathbf{F}^T\mathbf{V}_m\mathbf{F}| \right) - 2\log |\mathbf{F}^T\mathbf{F}| \right) \quad (9)$$

where the inner minimization is performed by enumerating over all possible \mathbf{k} and the outer using a simple coordinate exchange algorithm (Meyer and Nachtsheim, 1995).

4.3 Proof of concept

Suppose $n = 10$, and there are $q = 2$ inputs with $\mathbf{X} = [-1, 1]^2$, $\mathbf{f}^T(\mathbf{x}_i) = [1, x_{i,1}, x_{i,2}, x_{i,1}x_{i,2}, x_{i,1}^2, x_{i,2}^2]$ so that $d = 6$, and there is $g = 1$ PRN stream. The discrete set of values in the coordinate exchange algorithm is $\{-1, -0.9, -0.8, \dots, 0.8, 0.9, 1\}$. A robust design (denoted $(\mathbf{X}_R, \mathbf{k}_R)$) is found via (9)

where the prior joint distribution for ρ_- and ρ_+ used were independent uniform distributions. This is compared to the design (denoted \mathbf{X}_C) found by minimizing (9) but where the same PRN stream is used for all design points (i.e., CRN where $k(i) = 1, i = 1, \dots, n$), and to the design (denoted \mathbf{X}_I) found by minimizing (9) but with a different PRN stream for each design point (i.e., independent, which is equivalent to minimizing $-\log|\mathbf{F}^T\mathbf{F}|$ and is the standard D -optimal design). Note that these are the same comparisons as made by Schruben and Margolin (1978).

Both the independent and CRN designs converged to the face-centered central composite design with center point, while the robust design had repeated points at $[-1, -1]$ and $[+1, +1]$ and did not utilise the center point. The minimum values of $\log|\text{Var}(\hat{\boldsymbol{\beta}})|$ are $-9.1, -11.8$ and -13.0 for the independent, CRN and robust designs, respectively, thus demonstrating the benefit of inducing correlation over favouring independence and of the benefit of using a combination of common and antithetic random number streams.

Finally, an interesting comparison is the performance of the independent and CRN designs under the optimal allocation of the two PRN streams (R_1 and \bar{R}_1). It turns out that $\log|\text{Var}(\hat{\boldsymbol{\beta}})| = -11.8 > -13.0$ in both cases. This demonstrates the utility of jointly optimizing over the design points \mathbf{X} and PRN assignment \mathbf{k} , i.e. simply using a standard D -optimal design and then applying the optimal PRN assignment strategy to that design (as originally performed by Schruben and Margolin (1978)) can be outperformed by joint optimization.

5 Summary

Kleijnen (2015) provides important guidance on how to analyse simulation experiments in the event of departures from the ubiquitous independent and identically distributed assumptions. Heteroscedasticity in simulation output is not uncommon, and it is potentially beneficial to induce dependence through the reuse of pseudo-random number streams to reduce the generalized variance of the meta-model parameter estimators.

In this paper, we focus on the experimental design (vice analysis) aspects, and employed a computational approach to robust design for expensive computer experiments without the need to assume independence or identical distribution of errors in the meta-model to be developed. Through explicit modelling of the variance component for linear meta-models, the Fisher information was obtained within a maximum likelihood inference framework, while explicit modelling of the correlation structure for generalized linear meta-models and generalized estimating equations can be employed to approximate the Fisher information matrix. In both cases, robust designs can then be computationally sought which maximize some relevant statistic of this matrix, averaged across a prior distribution of any unknown parameters.

Moving away from the assumption of independence implies that a correlation structure be introduced, the misspecification of which could have a negative effect on the performance of the design. We built upon Woods and van de Ven (2011) to begin investigation of robust designs for GLMs with correlations. However, our work is distinct to Woods and van de Ven (2011) as we investigated the effect of covariance matrix misspecification for a variety of correlation structures, in the context of a 3-factor logistic GLM with pairwise interactions. While our results are not exhaustive, the cases of constant correlation, auto-regressive correlation, and distance correlation represent major classes of correlation structures (uniform, temporal, spatial) and are helpful to inform some recommendations.

As illustrated in Section 4, it may be effective to consider ρ as part of the vector of unknown parameters and hence integrate over a joint prior probability density. The choice to look at the efficiencies of designs as a function of ρ was primary to identify any dependencies between the effect of misspecification and the correlation strength. Since we mainly observe this dependency for the auto-regressive cases, robust design over ρ may only be required if auto-regressive is a feasible correlation structure. It is also important to note that this simulation approach is designed to obtain some

heuristics for dealing with correlation assumptions. In practice, the D -efficiency for a real problem will never be available. However, the results provide some means to assist in the interpretation of confidence regions that are obtained for a design. That is, one must assume some misspecification and therefore treat predicted parameter uncertainty estimates as underestimates for the true uncertainty that could arise when the computer experiment is performed.

Finally, Schruben and Margolin (1978) pioneered the search for effective assignment strategies of pseudo-random numbers to design points, but did so with fixed (textbook) designs (and for linear meta-models only). In this paper, we provide an example proof of concept of the possibility of jointly optimizing the design and pseudo-random number assignment and show that gains in statistical efficiency can be made.

6 Future research

This paper has assumed the vector of covariates, representing the simulator inputs and configuration settings, are continuous. However, discrete covariates must also be dealt with. Challenges arise in this case since the structure of covariances can be more complex. Furthermore, stochastic optimization is substantially more challenging to deal with in the discrete covariate case. While simulated annealing can deal with discrete spaces (Kirkpatrick et al., 1983), it will be more computationally intensive. Unfortunately, methods like Approximate Coordinate Exchange (Overstall and Woods, 2017) can only deal with continuous design spaces, however, other methods may exist to handle a discrete design space (Meyer and Nachtsheim, 1995). Further work is needed to determine the most effective computational scheme for this case.

Model misspecification is a broad challenge in robust design for computer experiments. While accounting for heteroscedasticity and correlations improves the situation substantially, there is still the potential for bias in the design due to the meta-model being unable to replicate some behaviours of the complex computer model. One approach to deal with this is the inclusion of an additional discrepancy term using Gaussian processes (Englezou, 2018; Kennedy and O’Hagan, 2000).

For the joint optimization of the design and PRN assignment, the coordinate exchange algorithm used relied on complete enumeration of all possible $\mathbf{k}(\cdot)$ assignments. The size of this set grows fast with the number of PRN streams g and would appear difficult to apply even for $g > 1$ and is therefore not particularly scalable. A more sophisticated approach will be required.

For non-linear meta-models we have focused on the logistic GLM that corresponds to a binary outcome from a simulation. However, the computational approach we consider here for robust design would also be applicable to other GLMs of interest, such as binomial and Poisson responses. For the joint optimization problem, a generalized linear mixed model (GLMM) approach might be applicable (see Chapter 17 of Pawitan (2013)). The meta-model would then have the form $g(\mathbb{E}_{\mathbf{Y}}[\mathbf{Y}|\mathbf{X}]) = \mathbf{F}\boldsymbol{\beta} + \boldsymbol{\gamma}$, where $\boldsymbol{\gamma} = \mathbf{Z}\mathbf{b}$ as in Section 4. The unknown parameters $\boldsymbol{\beta}$ could be estimated via maximum likelihood where $\text{Var}(\hat{\boldsymbol{\beta}}) \approx (\mathbf{F}^T \mathbf{M} \mathbf{F})^{-1}$ and there are various forms for \mathbf{M} under different approximations (Pawitan, 2013). This modelling approach is an example of a GLMM, for which optimal design have been considered previously (see Xu and Singh (2021) and references therein).

References

- Atkinson, A. C., and R. D. Cook. 1995. “ D -Optimum Designs for Heteroscedastic Linear Models”. *Journal of the American Statistical Association* 90:204–212.
- Chaloner, K., and I. Verdinelli. 1995. “Bayesian Experimental Design: A Review”. *Statistical Science* 10(3):273–304.

- Donev, A., A. Atkinson, and R. Tobias. 2007. *Optimum Experimental Designs, with SAS*. Oxford Statistical Science Series. United Kingdom: Oxford University Press.
- Dunn, P., and G. Smyth. 2018. *Generalized Linear Models with Examples in R*. Springer Texts in Statistics. Springer New York.
- Englezou, Y. 2018, July. *Bayesian Design for Calibration of Physical Models*. Ph. D. thesis, University of Southampton. <https://eprints.soton.ac.uk/427145/>, accessed 6th May 2021
- Fedorov, V. V. 1972. *Theory of Optimal Experiments [by] V.V. Fedorov. Translated and edited by W.J. Studden and E.M. Klimko*. Academic Press New York.
- Gill, A. 2019. “Two Common Pitfalls Applying Design of Experiments (and Hopefully How to Avoid Them!)”. In *Proceedings of MODSIM2019, 23rd International Conference on Modelling and Simulation*, edited by S. Elsworth, 323–329. Canberra, Australian Capital Territory, Australia: Modelling and Simulation Society of Australia and New Zealand, Inc.
- Gill, A. 2021. “Heteroscedasticity and Correlation in Linear Regression”. In *Proceedings of MODSIM2021, 24th International Conference on Modelling and Simulation*, edited by R. W. Vervoort, A. A. Voinov, J. P. Evans and L. Marshall, 834–840. Canberra, Australian Capital Territory, Australia: Modelling and Simulation Society of Australia and New Zealand, Inc.
- Gill, A., D. Grieger, M. Wong, and W. Chau. 2018. “Combat Simulation Analytics: Regression Analysis, Multiple Comparisons and Ranking Sensitivity”. In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, S. Jain and B. Johansson, 3789–3800. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.
- Goos, P., and B. Jones. 2011. *Optimal Design of Experiments: A Case Study Approach*. Wiley.
- Kennedy, M. C., and A. O’Hagan. 2000. “Bayesian Calibration of Computer Models”. *Journal of the Royal Statistical Society, Series B, Methodological* 63:425–464.
- Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. 1983. “Optimization by Simulated Annealing”. *Science* 220(4598):671–680.
- Kleijnen, J. 2015. *Design and Analysis of Simulation Experiments*. 2nd ed. New York, USA: Springer.
- Levenberg, K. 1944. “A Method for the Solution of Certain Non-Linear Problems in Least Squares”. *Quarterly of Applied Mathematics* 2:164–168.
- Liang, K.-Y., and S. L. Zeger. 1986, 04. “Longitudinal Data Analysis Using Generalized Linear Models”. *Biometrika* 73(1):13–22.
- Marquardt, D. W. 1963. “An Algorithm for Least-Squares Estimation of Nonlinear Parameters”. *SIAM Journal on Applied Mathematics* 11(2):431–441.
- Meyer, R., and C. Nachtsheim. 1995. “The Coordinate-Exchange Algorithm for Constructing exact Optimal Experimental Designs”. *Technometrics* 37:60–69.
- Montgomery, D. 2012. *Design and Analysis of Experiments, 8th Edition*. John Wiley & Sons, Incorporated.
- Overstall, A. M., and D. C. Woods. 2017. “Bayesian Design of Experiments using Approximate Coordinate Exchange”. *Technometrics* 59(4):458–470.

- Pawitan, Y. 2013. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press.
- Sanchez, S. M., T. W. Lucas, P. J. Sanchez, C. J. Nannini, and H. Wan. 2012. *Designs for Large-Scale Simulation Experiments, with Applications to Defense and Homeland Security*, Chapter 12, 413–441. John Wiley & Sons, Ltd.
- Santner, T. J., W. B., and N. W.. 2003. *The Design and Analysis of Computer Experiments*. New York, USA: Springer-Verlag.
- Schruben, L. W., and B. H. Margolin. 1978. “Pseudorandom Number Assignment in Statistically Designed Simulation and Distribution Sampling Experiments”. *Journal of the American Statistical Association* 73(363):504–520.
- Weiser, C. 2016. *mvQuad: Methods for Multivariate Quadrature*. (R package version 1.0-6). <https://cran.r-project.org/web/packages/mvQuad/index.html>, accessed 13th April 2022.
- Woods, D. C., S. M. Lewis, J. A. Eccleston, and K. G. Russell. 2006. “Designs for Generalized Linear Models with Several Variables and Model Uncertainty”. *Technometrics* 48:284–292.
- Woods, D. C., and P. van de Ven. 2011. “Blocked Designs for Experiments With Correlated Non-Normal Response”. *Technometrics* 53(2):173–182.
- Xu, X., and S. Singh. 2021. “Robust Designs for Generalized Linear Mixed Models with Possible Model Misspecification”. *Journal of Statistical Planning and Inference* 210:20–41.