Oral examination: Optimal Design for Nonlinear Models

Ryan Jarrett

May 1, 2017

1 Introduction

In this paper we explore some of the work that has been done on the optimal design of experiments. In general, we wish to answer the following question: "given a pre-specified model and an estimator of interest, where should we take measurements so as to best estimate parameters of interest with the most precision?" As we will discuss later, what qualifies as "optimal" largely depends on the investigator's interests. A design that may be considered optimal for a specific purpose, such as minimizing the variance of a parameter estimate, may not be optimal for another purpose, such as minimizing the correlation between two parameters that we wish to jointly estimate. The theory discussed here assumes that one is interested in model-based inference. This paper and the examples contained focus on non-linear models, though the same theory holds for linear models as a special case. In particular, we consider an application to a pharmacokinetic (PK) model in which we are interested in identifying the most informative times at which to draw a patient's blood, so as to estimate the concentration of a medication within their blood stream as a function of time. The majority of the theory presented in this paper draws from Valerii V. Fedorov and Sergei L. Leonov's 2014 textbook "Optimal Design for Nonlinear Response Models." Additional sources are used primarily to supplement Fedorov & Leonov's work or to explore related topics in the field of optimal design.

2 Estimators and Notation

We consider models of the form

$$\begin{aligned} y_{ij} &= \eta(x_i, \theta) + \epsilon_{ij} \\ \epsilon_{ij} &\sim N(0, \sigma^2) \end{aligned} \tag{1}$$

for $i = 1, 2, ..., n$ unique design points and $j = 1, ..., r_i$ replications at each design point, such that $\sum_{i=1}^{n} r_i = N$. That is, y is some function of x, parameterized by θ, with independent normal errors that have mean 0 and variance σ^2. Our goal is to identify a set of design points, x, that can be considered "optimal" according to some prespecified criterion. For all examples considered, this criterion will be a function of the information matrix, defined and discussed extensively in the following text. We do not require $\eta(x, \theta)$ to be linear in x or theta. In the subsequent definitions, we express formula as functions of both, x and θ to allow for the more general non-linear case in which this is true. For linear models (i.e. models that are linear in θ, such that they can be expressed in the form $\eta(x, \theta) = X\theta$), the exact same theory will hold with the added simplification many terms are functions only of x, and therefore do not rely on the unknown parameter values, θ. We also define the "experiment", ξ_N, as the set of design points and its replicates:

$$\xi_N = \{x_i, p_i\}_1^n \tag{2}$$

where $p_i = r_i/N$. In the subsequent theory, in many cases to identify an optimal design we will allow p_i to be continuous on the range $(0, 1)$ and convert back to a whole number of replications at each design point by multiplying $r_i = Np_i$. The gradient vector with respect to θ and the information matrix play a central role within the following theory and are defined as follows:

$$\frac{\partial \eta(x, \theta)}{\partial \theta} = \nabla \eta \tag{3}$$

We will denote the $N \times k$ dimensional gradient as $\nabla \eta_N$, where N is the total number of observations and k is the dimensionality of θ. The information matrix with respect to a single observation is

$$\mu(x, \theta) = \sigma^{-2} \nabla \eta \nabla \eta^T \tag{4}$$

The information matix is additive for i.i.d. observations, therefore the full information matrix for all N observations is given as

$$\underline{\mathrm{M}}(\xi_N, \theta) = \sigma^{-2} \sum_{i=1}^{n} r_i \mu(x_i, \theta) \tag{5}$$

We define $\underline{\mathrm{D}}(\xi_N, \theta)$ to be the inverse information matrix.

$$\underline{\mathrm{D}}(\xi_N, \theta) = \underline{\mathrm{M}}(\xi_N, \theta)^{-1} \tag{6}$$

In the linear case, this is exactly equal to the variance-covariance matrix, while in non-linear cases it is an approximation to the variance-covariance matrix, with the degree of accuracy depending on the closeness of the Taylor approximation of $\eta(x, \theta)$ about θ to the true function. It is useful in many cases to distinguish between the "Information Matrix" and the "Fisher Information Matrix," where the former is defined as above, while the latter is defined as the negative expectation of the second derivative of the log likelihood:

$$I(\theta) = -E_\theta \left[\frac{\partial^2 \ln\, p(x|\theta)}{\partial \theta \partial \theta^T} \right] = -E_\theta[H(\theta)]$$

For linear models with normal errors, however, the information and Fisher information matrices coincide with the value $\sigma^{-2} X^T X$. In the non-linear case with normal errors, the formula provided in equation 5 provides an approximation to the Fisher Information Matrix based on a first-order Taylor series expansion. Derivations of the Fisher information matrices for the linear and nonlinear cases are provided in the Appendix. We also note that the information matrix is symmetric, non-negative definite matrix (i.e. all of its eigenvalues are greater than or equal 0),

$$z^T \underline{\mathrm{M}}(\xi_N, \theta) z = \sigma^{-2} \sum_i r_i [z^T \nabla \eta]^2 \geq 0 \tag{7}$$

for any vector $z \in \mathbb{R}^m$. It is also useful to introduce a normalized information matrix for all observations:

$$M(\xi_N, \theta) = \int_{\mathcal{X}} \mu(x, \theta_t) \xi(dx) \tag{8}$$

Where $\xi(dx)$ is the design weight given at value x. In the case where all design points are given equal weight, we have $M(\xi_N, \theta) = \underline{\mathrm{M}}(\xi_N, \theta)/N$.

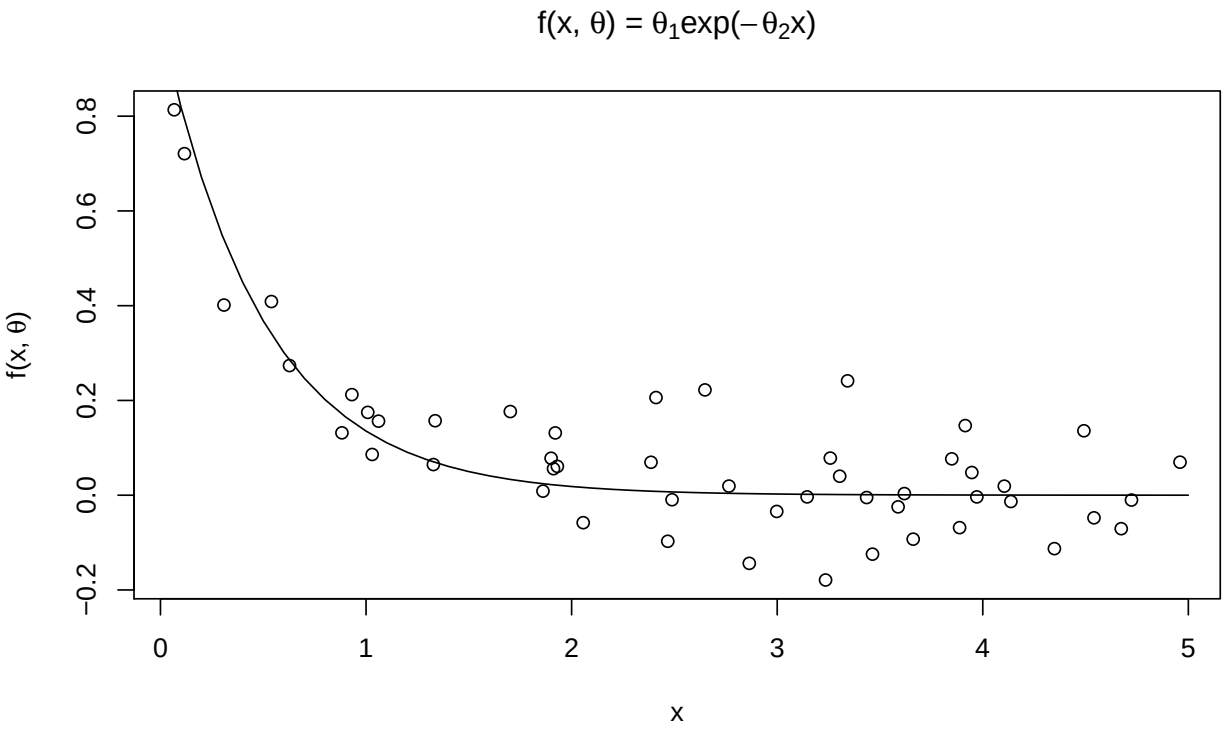


Figure 1: Non-linear regression fit to generated data evaluated at the true values of theta.

3 Geometric Interpretation

All of the optimality criteria defined later in the paper are expressed as functions of the information matrix. Consequently, we believe it helpful to consider some of the geometric characteristics of the information matrix. For illustration, we consider the simple nonlinear model

$$y_i = \theta_1 e^{-\theta_2 x_i} + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

with values $\{\theta_1 = 1, \theta_2 = 2, \sigma = 0.1, N = 50\}$. Our gradient vector is given

$$\nabla \eta^T = (e^{-\theta_2 x}, -\theta_1 x e^{-\theta_2 x})$$

and information matrix for N observations

$$\underline{\mathrm{M}}(\xi_N, \theta) = \sigma^{-2} \nabla \eta_N^T \nabla \eta_N$$

Where η_N is a $N \times 2$ matrix of gradient functions evaluated at each observation, x. The first-order Taylor series expansion of $\eta(x, \theta)$ about θ_t is given

$$\eta(x, \theta) = \eta(x, \theta_t) + (\theta - \theta_t) \nabla \eta$$

We sample 50 random values of x uniformly on the bounds $[0, 5]$ and generate corresponding values of y using the mean model and error described above. Using an iterative algorithm we calculate the least-squares estimates, $(\hat{\theta}_1, \hat{\theta}_2) = (0.87, 1.67)$. The generated data are shown in Figure 1 with the line corresponding to the true model fit overlaid. The covariance matrix for the $\hat{\theta}$ estimates is

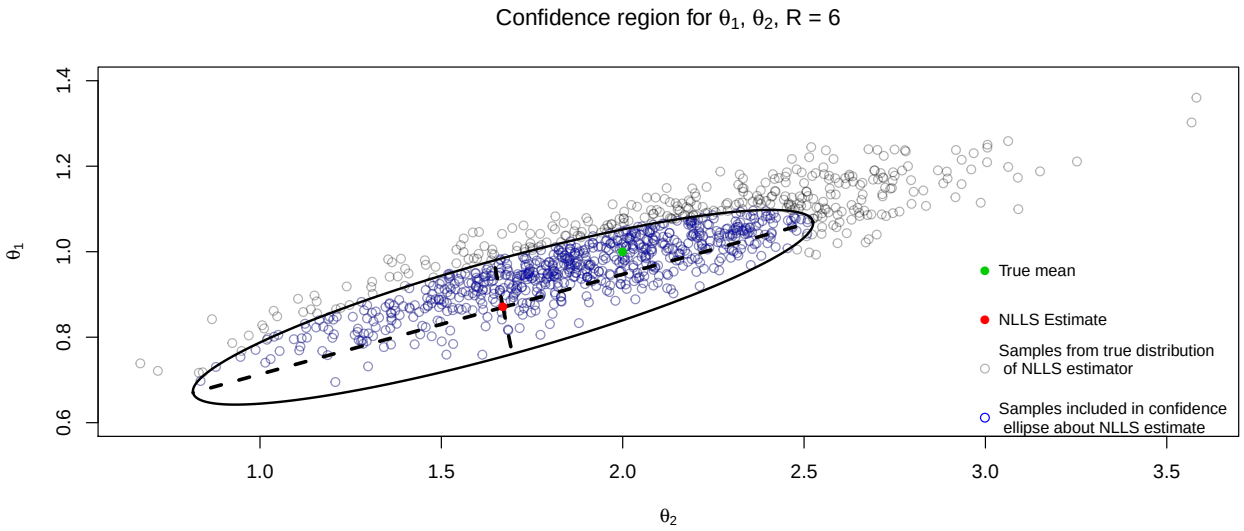


Figure 2: Confidence ellipse corresponding to non-linear least-squares (NLLS) estimates of theta at the 0.95 significance level. Points represent samples from the true sampling distribution of the estimator of θ. Each axis of the confidence ellipse represents a principal direction of the variance covariance matrix.

approximated by the inverse of the information matrix and can be decomposed into its eigenvectors and eigenvalues as follows:

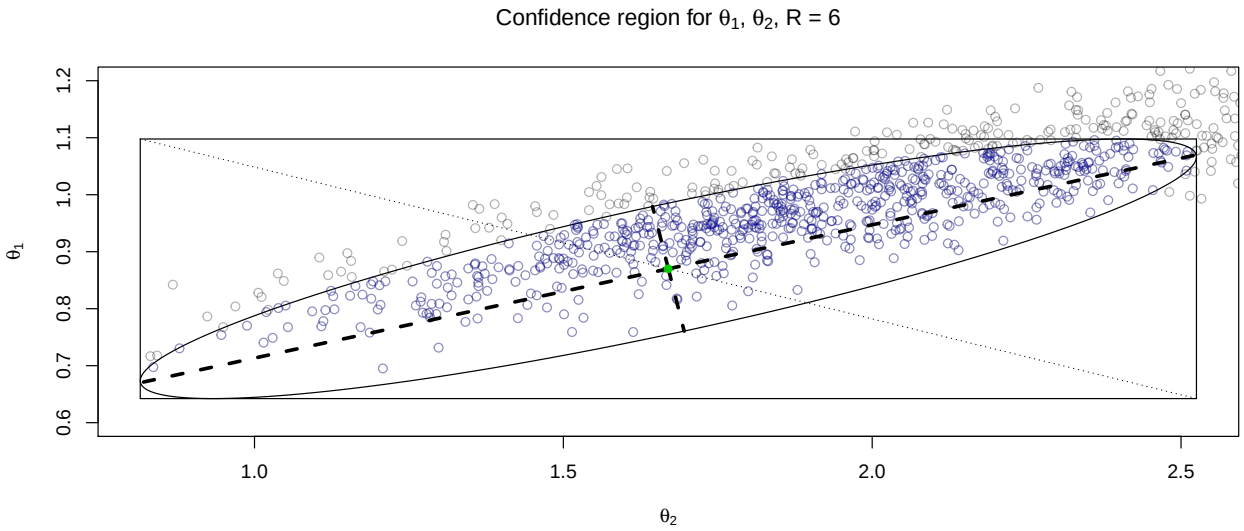
$$\underline{\mathrm{D}} = V \Lambda V^T$$

Where V is an orthonormal matrix of eigenvectors and Λ is a diagonal matrix of eigenvalues. The eigenvectors, V, define the principal axes of the confidence ellipse for θ, while the eigenvalues indicate the amount of variation along each axis. It is worth noting that since $\underline{\mathrm{D}}^T \underline{\mathrm{M}} = I$, the eigenvectors of $\underline{\mathrm{D}}$ are identical to those of $\underline{\mathrm{M}}$, while the eigenvalues are reciprocals of each other. Consequently, when we modify a design to increase the eigenvalue corresponding to a principal axis of $\underline{\mathrm{M}}$, we proportionally decrease the variance of $\hat{\theta}$ along the same axis. In general, we do this by increasing the variation in the gradient matrix, $\nabla\eta(x, \theta)$ (i.e. we increase the sum of the squared deviations about a mean). For linear models in which $\eta(x, \theta) = X\theta$ and $\nabla\eta(x, \theta) = X$, this corresponds to increasing the variability in X.

The criteria for optimality that follow are defined in terms of the eigenvalues of the information matrix or its reciprocal. Consequently, it is useful to keep these properties in mind so as to recognize that each criterion has a distinct and intuitive geometric interpretation (e.g. maximizing the volume of the data ellipse, maximizing the length of the shortest axis, etc.). In all cases explored within this paper, we are concerned with some aspect of the shape of an ellipsoid of concentration, given by the the quadratic formula

$$CI_{1-\alpha,\ \chi^2} = \left\{ \theta : (\theta - \hat{\theta}_N)^T \underline{\mathrm{M}}(\xi_N, \hat{\theta}_N)(\theta - \hat{\theta}_N) \leq \chi^2_{k,\alpha} \right\} \tag{9}$$

where k is the dimension of θ. Figure 2 shows the ellipsoid of concentration corresponding to the non-linear regression model detailed above. While the square-root of the eigenvalues of $\underline{\mathrm{D}}$ define the relative half-lengths of the principal axes shown in Figure 2, to calculate a $100 * (1 - \alpha)\%$ confidence region, we must scale each axis by the square-root of the χ^2 critical value. For example, the 95th percentile of a χ^2 distribution with $k = 2$ degrees of freedom for the two parameters estimated is $R = 5.99 \approx 6$. The half-lengths of the principal axes for the ellipse shown in Figure 2 are equal to $(\sqrt{\lambda_1 R}, \sqrt{\lambda_2 R}) = (\sqrt{(0.129)(5.99)}, \sqrt{(0.002)(5.99)}) = (0.878, 0.112)$, respectively.



4 Optimality Criteria

How do we define what constitutes an optimal design? Since the variance of our estimates is either given or approximated by the inverse of the information matrix. Naturally, we may want consider a definition of optimality of the form

$$\xi_N^* = \arg\max_{\xi_N} \underline{\mathrm{M}}(\xi_{\mathbf{N}}, \theta) = \arg\min_{\xi_N} \underline{\mathrm{D}}(\xi_{\mathbf{N}}, \theta) \tag{10}$$

as such a criteria would minimize the variance of our estimates, $\hat{\theta}$. Since $\underline{\mathrm{M}}(\xi_{\mathbf{N}})$, $\underline{\mathrm{D}}(\xi_{\mathbf{N}})$ are matrices, we could understand maximizing or minimizing them in terms of a Loewner order, such that $\underline{\mathrm{M}}(\xi_{\mathbf{N}}) > \underline{\mathrm{M}}(\xi'_{\mathbf{N}})$ if $\underline{\mathrm{M}}(\xi_{\mathbf{N}}) - \underline{\mathrm{M}}(\xi'_{\mathbf{N}})$ is positive definite, and $\underline{\mathrm{M}}(\xi_{\mathbf{N}}) \geq \underline{\mathrm{M}}(\xi'_{\mathbf{N}})$ if $\underline{\mathrm{M}}(\xi_{\mathbf{N}}) - \underline{\mathrm{M}}(\xi'_{\mathbf{N}})$ is positive semi-definite. Intuitively, such a criteria would mean that the information matrix corresponding to an optimal design has as much or more variation along all of its principal axes than those of all other sub-optimal designs. Fedorov and Leonov (2014), however, prove that even in the case of a simple linear regression, there does not exist a design that meets such a stringent optimality criterion. Instead, we are required to consider the minimization of a function, Ψ, of the information matrix:

$$\xi_N^* = \arg\min_{\xi_N} \Psi[\underline{\mathrm{M}}(\xi_N, \theta)] \tag{11}$$

4.1 Commonly used criteria

Commonly used optimality criteria include the following:

D-Optimality

$$\Psi = |\underline{\mathrm{D}}(\xi_N, \theta)| = |\underline{\mathrm{M}}(\xi_N, \theta)|^{-1} \tag{12}$$

The D-optimality criterion minimizes the determinant of the variance-covariance matrix for $\hat{\theta}$ and, in doing so, minimizes the area/volume of the the confidence ellipse. The determinant of the variance-covariance matrix is given by the product of the eigenvalues of $\underline{\mathrm{D}}(\xi_N, \theta)$. The volume/area of a data ellipse is proportional (proportionality constant π in 2 dimensions, $\frac{4\pi}{3}$ in 3 dimensions) to the product

of the root eigenvalues. For an arbitrary number of dimensions, the volume of the confidence ellipsoid is proportional to $|\underline{\mathrm{D}}(\xi_{\mathbf{N}}, \theta)|^{1/2}$ and is given by the formula

$$\text{Volume} = \frac{\pi^{k/2}}{\Gamma(k/2+1)} R^{k/2} |\underline{\mathrm{D}}(\xi_{\mathbf{N}}, \theta)|^{1/2}$$

Where k is the dimensionality of θ and R is the bound of the ellipsoid, given by the χ^2 critical value associated with the confidence ellipse in equation 9.

E-Optimality

$$\Psi = \lambda_{\min}^{-1} |\underline{\mathrm{M}}(\xi_{\mathbf{N}}, \theta)| = \lambda_{\max} |\underline{\mathrm{D}}(\xi_{\mathbf{N}}, \theta)| \tag{13}$$

In choosing the E-Optimality criterion, we minimize the largest eigenvalue of the variance-covariance matrix. In doing so, we aim to shorten the longest axis length of our confidence ellipse, thereby minimizing the largest variance associated with any linear combination of the parameter estimates.

A-Optimality

$$\Psi = \mathrm{tr}|A\underline{\mathrm{D}}(\xi_N, \theta)| \tag{14}$$

Where A is a nonnegative definite matrix, referred to as a utility matrix. When A is a diagonal matrix with non-negative elements, the A-Optimality criterion minimizes the weighted average variance of the parameter estimates with the jth weight given by the jth diagonal element of A divided by the trace of A. In the case where A is the identity matrix, this geometrically corresponds to minimizing the volume of the k-dimensional hyperrectangle (in our example, the 2-dimensional rectangle) that contains the confidence ellipse.

We present these criteria primarily to illustrate that there exist many properties of the confidence ellipse that we may wish to optimize, and that an "optimal" design depends on which properties of an estimator one is interested in. For the sake of succinctness, the remainder of the paper will focus exclusively on the D-criterion, as it is the most widely used. However, the approach to identifying an optimal design is similar for other optimality criteria.

4.2 Approaches to parameter uncertainty

For nonlinear models the optimality criteria is a function of the unknown parameters, θ. We must, therefore, find some way to account for this uncertainty. The simplest approach is referred to as a "locally optimal" design, in which one plugs in a guess, θ_0, for θ. While straightforward, this approach falls somewhat short of inspiring confidence in the resulting design if the design points are sensitive to the values of θ and we are not certain of our chosen values of θ. Another approach is to use a "minimax" design, in which one finds the design that minimizes the design criterion at the parameter value that maximizes the criterion. In this sense, we find the best design at the worst possible values of θ.

$$\xi_N^* = \underset{\xi_N}{\operatorname{argmin}} \max_{\theta \in \Theta} \Psi[\underline{\mathrm{M}}(\xi_N, \theta)] \tag{15}$$

Another approach is to minimize the average criterion value by integrating over an assumed distribution for the unknown parameter:

$$\xi_N^* = \underset{\xi_N}{\operatorname{argmin}} \int_\Theta \Psi[\underline{\mathrm{M}}(\xi_N, \theta)]\mathcal{A}(d\theta) \tag{16}$$

where $\mathcal{A}(d\theta)$ is a prior distribution on θ.

It is also possible to adaptively estimate θ based on sequential sampling (Fedorov & Leonov, 2014). The general approach is to guess at θ_0 and sample according to the optimal design under θ_0, estimate $\hat{\theta}$ from the first set of experiments, and update the optimal design based on $\hat{\theta}$. We consider later a sampling approach, suggested by Matt Shotwell, in which the optimal design points are identified for a set of parameter values, drawn randomly from the prior distribution. The primary benefit of this approach is that it allows one to identify the design points that are robust to chosen parameter values, and illustrates the degree of variablilty in the design points that are not robust to parameter values.

4.3 Necessary and sufficient conditions for optimality

For a specific design to be optimized, it is necessary and sufficient for it to meet the following four criteria (Fedorov & Leonov, 2014):

1. $\Psi(M)$ is a convex function, where covexity is defined such that if $M = (1-\alpha)M_1 + \alpha M_2$, then

$$\Psi(M) \leq (1-\alpha)\Psi(M_1) + \alpha\Psi(M_2) \tag{17}$$

2. $\Psi(\underline{\mathrm{M}})$ is a monotonically non-increasing function, such that if $M \geq M'$ in terms of Lowener ordering, then $\Psi(\underline{\mathrm{M}}) \leq \Psi(\underline{\mathrm{M}}')$.

To reiterate, $\underline{\mathrm{M}} \geq \underline{\mathrm{M}}'$ in Lowener ordering if the eigenvalues of $\underline{\mathrm{M}} - \underline{\mathrm{M}}' \geq 0$. In practice, due to the additivity of the information matrix, $\underline{\mathrm{M}}$ becomes larger in Lowener ordering as additional observations are made. Consequently, as N increases, $\underline{\mathrm{M}}$ never decreases in terms of Loewner ordering.

3. Let $\Xi(q) = \{\xi : \Psi|M(\xi)| \leq q < \infty\}$ Then there exists a real number q such that $\Xi(q)$ is non-empty.

This assumption merely requires that designs considered have a finite value of the chosen optimality criterion.

4. For any $\xi, \bar{\xi} \in \Xi(q)$

$$\Psi_\alpha(\xi, \bar{\xi}) = \Psi[(1-\alpha)M(\xi) + \alpha M(\bar{\xi})] = \Psi[M(\xi)] + \alpha \int_{\mathcal{X}} \psi(x, \xi)\bar{\xi}(dx) + o(\alpha|\xi, \bar{\xi}) \tag{18}$$

where $\lim_{\alpha \to 0} \frac{o(\alpha|\xi,\bar{\xi})}{\alpha} = 0$ and $0 \leq \alpha \leq 1$.

In plain terms, this assumption states that for any two designs, ξ and $\bar{\xi}$, the value of the optimality criterion associated with a weighted sum of the two corresponding information matrices can be expressed as the optimality criterion of one design plus a weighted measure of distance or dissimilarity between the two designs. The amount of weight is given by α, while the "distance" is what is called a directional (Gâteaux) derivative (plus a remainder that tends to 0 as α becomes smaller). The assumption requires that the directional derivative with regard to the optimality criterion function takes the form

$$\frac{\partial \Psi_\alpha(\xi, \bar{\xi})}{\partial \alpha} = \int_{\mathcal{X}} \psi(x, \xi)\bar{\xi}(dx) \tag{19}$$

Intuitively, the directional derivative (i.e. the terms after α) tells us how much of a change in the optimality criterion value is associated with a movement from design ξ to design $\bar{\xi}$ across the entire support of $\mathcal{X}$, while α tells us the magnitude of the change. Under the integral, $\psi(x, \xi)$ tells us the magnitude of the change in $\Psi(\xi)$ associated with a small increase in weight at point x and a corresponding decrease in weight at other values. $\bar{\xi}(dx)$ provides the weight given to point x by the design $\bar{\xi}$. By integrating the change at each value, x, times the weight given to point x by design $\bar{\xi}$ over the support $\mathcal{X}$, we calculate the approximate change in the optimality criterion associated with an $\alpha-$weighted shift from design ξ to design $\bar{\xi}$.

For each of the common criteria, the directional derivatives have known forms and are merely presented here, though a derivation for the D-criterion is given in the Appendix.

$$\psi^D(x, \xi, \theta) = m - \text{tr}[M^{-1}(\xi, \theta)\mu(x, \theta)] \tag{20}$$

where m is the number of estimated parameters and $\mu(x)$ is the information matrix for a single observation, given by equation 4. The sensitivity function for the D-criterion is defined as

$$d(x, \xi, \theta) = -\psi^D(x, \xi, \theta) + m = \text{tr}[M^{-1}(\xi, \theta)\mu(x, \theta)] \tag{21}$$

For E and A optimality, it is given by

$$\varphi(x, \xi, \theta) = -\psi(x, \xi, \theta) + \Psi(\xi, \theta) \tag{22}$$

The sensitivity function plays a large role in the search for and identification of optimal designs, detailed below.

5 Identification of an Optimal Design

A variety of methods can be used to identify an optimal design. If the number of design points, N, is sufficiently small and the support of X is discrete (or can be discretized), then a brute-force calculation of the optimality criterion at all combinations could be used to find the design that minimizes the optimality criterion. In many cases, however, this approach is infeasible. When N is large but the sample space is discrete, various search algorithms can be used to find an optimal design.

5.1 Equivalence theorems

For the optimality criteria listed above, there exist "Equivalence Theorems," which express optimality criteria in alternate forms that can quickly verify whether or not a design is optimal. For example, for the D-criterion, the following criteria were shown to be equivalent (Kiefer-Wolfowitz, 1960):

$$\begin{aligned} &1.\ \min_\xi |D(\xi, \theta)| \\ &2.\ \min_\xi \max_x d(x, \xi) \\ &3.\ \max_x d(x, \xi) = m \end{aligned} \tag{23}$$

Where $d(x, \xi)$ is the sensitivity function for the D-optimality criterion, defined in equation 21. Similar equivalence theorems can also be stated for other commonly used criteria, but are omitted here. Primarily, we will concern ourselves with the first and third forms of the equivalence theorem, which state that the determinant of the variance-covariance matrix is minimized when the maximum of the sensitivity function on the support of x is equal to the number of parameters in the model, m. It is

worth noting, however, that $d(x, \xi, \theta)$ is approximately equal to the variance of the estimated response (equal in the linear case), and reaches its minimum at $d(x, \xi, \theta) = m$. Therefore, in finding a D-optimal design, in addition to minimizing the volume of the confidence ellipse around our estimated parameters, the maximum variance of the estimated response is also minimized. A proof of the equivalence of the sensitivity function and the variance of the predicted response is given in the appendix.

5.2 Algorithm for identification

Though faster algorithms are available, we choose a relatively simple procedure for finding an optimal design, as described by Federov and Leonov's forward and backward iterative procedures. The process consists of iteratively finding the points on the support of x at which the value of the sensitivity function is the highest and lowest. These points are then up-weighted and down-weighted, respectively in the next iteration by a scalar, α_s, that varies with each iteration.

For each iteration until convergence of $\Psi(M)$ within some tolerance we apply the following algorithm:

1. Given the design at the s^{th} iteration, ξ_s, find the points that maximize and minimize the sensitivity function, respectively:

$$x_{s+1}^{+} = \arg\max_{x} d(x, \xi_s)$$
$$x_{s+1}^{-} = \arg\min_{x} d(x, \xi_s)$$

2. Add the point x_{s+1}^{+} to the design with weight α_s^{+}.

$$\xi_s^{+} = (1 - \alpha_s^{+})\xi_s + \alpha_s^{+}\xi(x_{s+1}^{+})$$

   Fedorov & Leonov show the α that achieves the steepest descent (i.e. it minimizes the optimality criterion evaluated at the resulting design) to be

$$\alpha_s^{+} = \arg\min_{\alpha} |D(\xi_{s+1})| = \frac{d(x_{s+1}^{+}, \xi_s) - m}{[d(x_{s+1}^{+}, \xi_s) - 1]m}$$

3. Down-weight x_{s+1}^{-} in the design ξ_s^{+}

$$\xi_{s+1} = (1 - \alpha_s^{-})\xi_s^{+} + \alpha_s^{-}\xi(x_{s+1}^{+})$$

   Where

$$\alpha_s^{-} = -\min\left(\alpha_s^{+}, \frac{p_s}{1 - p_s}\right)$$

   and p_s is the weight given to point x_{s+1}^{-} in design ξ_s.

We initiate the algorithm with a design, ξ_0, that uniformly distributes sampling weight over the support of $\mathbf{x}$. As a practical consideration, while the design weight, ξ, is a continuous quantity, the support of $\mathbf{x}$ must be descretized at some level of precision; however, this is not a major concern in the identification of design points in practice. For example, if one wishes to identify the optimal time points at which to draw a patient's blood, it is likely sufficient to consider time as discretized into minute-long intervals. The only cost associated with discretizing to increasingly granular intervals is additional computing time.

6 Application to Pharmacokinetics Data

We now consider an application of these methods to a Pharmacokinetics (PK) example. Here, we are interested in identifying the times at which a patient's blood should be drawn to most accurately estimate the drug concentration over time, given a pre-specified dosing schedule. A two-compartment differential equation model is used to describe the dynamics of the drug within an individual. This model allows for a central compartment (i.e. bloodstream), with a mass of the drug given by m_1, and a peripheral compartment (i.e. tissue) with mass m_2. The two compartments have volumes v_1 and v_2, respectively. The rates of diffusion from the central to the peripheral is defined as k_{12}, and from the peripheral to the central as k_{21}. An elimination rate from the central component is given by the parameter k_{10}. The drug infusion rate is given by k_R and is assumed to be known, as this is controlled by the experimenter/physician. For the present example, k_R is a piecewise function of time, taking on a positive constant value while the drug is being infused intravenously, and 0 for all other times. The model can be expressed as a system of two equations relating the change in drug concentration within each compartment over time to the four unknown PK parameters, $\{v, k_{12}, k_{21}, k_{10}\}$.

$$\begin{aligned} \frac{dm_1}{dt} &= k_R + k_{21}m_2 - k_{12}m_1 - k_{10}m_1 \\ \frac{dm_2}{dt} &= k_{12}m_1 - k_{21}m_2 \end{aligned} \tag{24}$$

The mass in the two compartments can easily be converted to a concentration by dividing by the volumes of the respective compartments: $c_1 = m_1/v_1$ $c_2 = m_2/v_2$. In practice, we typically observe a measurement of c_1, the concentration of the drug within the main compartment (i.e. bloodstream). The solutions to this set of equations provide models of the total drug concentration in a patient's body as a function of time. A derivation of the solution is provided in the Appendix. We assume the measured drug concentration in the central compartment at time t to be normally distributed about the value predicted by the solution for $c_1(t)$ in the above differential equations with variance σ^2.

$$c_1(t) \sim N(g(t, v_1, k_{10}, k_{12}, k_{21}), \sigma^2) \tag{25}$$

Estimates of population PK parameters and σ^2 from a previous study are used to create prior distributions for the PK parameters and error variance, respectively. These prior distributions are, therefore, best thought of the estimated variability within the population, rather than our state of uncertainty regarding their true values, as we frequently use priors to represent. Specifically, we use the following prior distributions:

$$\begin{aligned} \{\ln v_1, \ln k_{10}, \ln k_{12}, \ln k_{21}\} &\sim N_4\left(\mu_0 = \begin{pmatrix} 3.223 \\ -1.650 \\ -5.000 \\ -5.000 \end{pmatrix}, \Sigma_0 = \begin{pmatrix} 0.501 & -0.384 & 0.000 & 0.000 \\ -0.384 & 0.462 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.045 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.045 \end{pmatrix}\right) \\ \ln(\sigma) &\sim N(2.33, 0.32) \end{aligned} \tag{26}$$

An example central concentration curve as a function of time for a fixed set of prior parameters with simulated observations is given in Figure 3. In the figure, PK parameter values and an error term are drawn from the prior distribution specified. This set of parameter values can be thought of as characterizing the pharacokinetic profile of a hypothetical patient within the population of patients described by the prior distribution. The patient then receives 5 intravenous doses, with each infusion lasting 3 hours. The dosing interval (i.e. the time between the starts of consecutive infusions) is set at 8 hours, with measurements are taken during the fifth dosing period at t = (32, 32.5, 33, 34, 36, 38) hours.

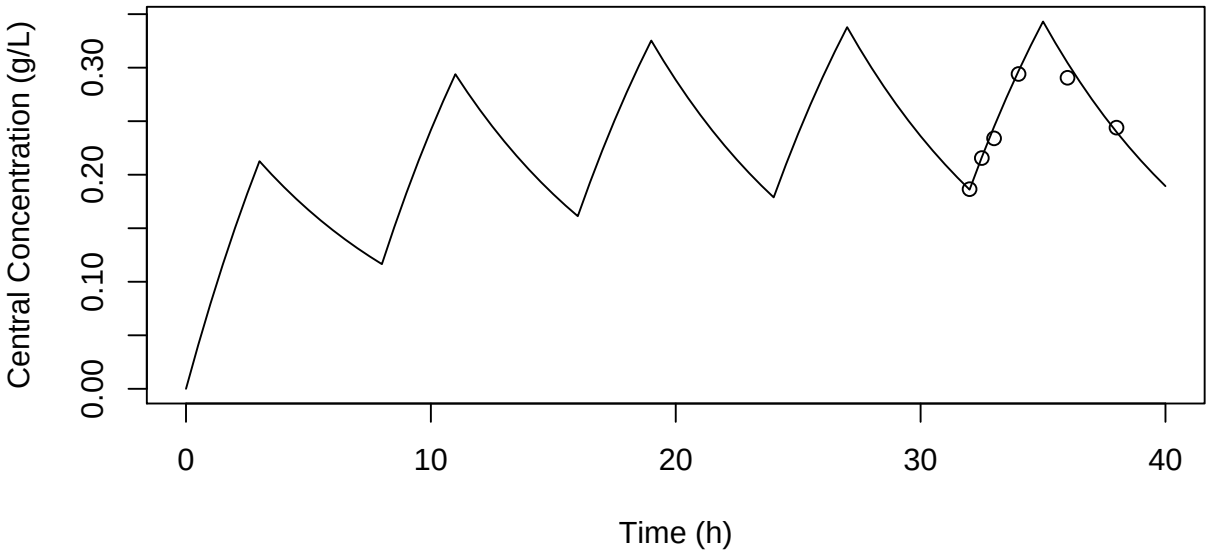
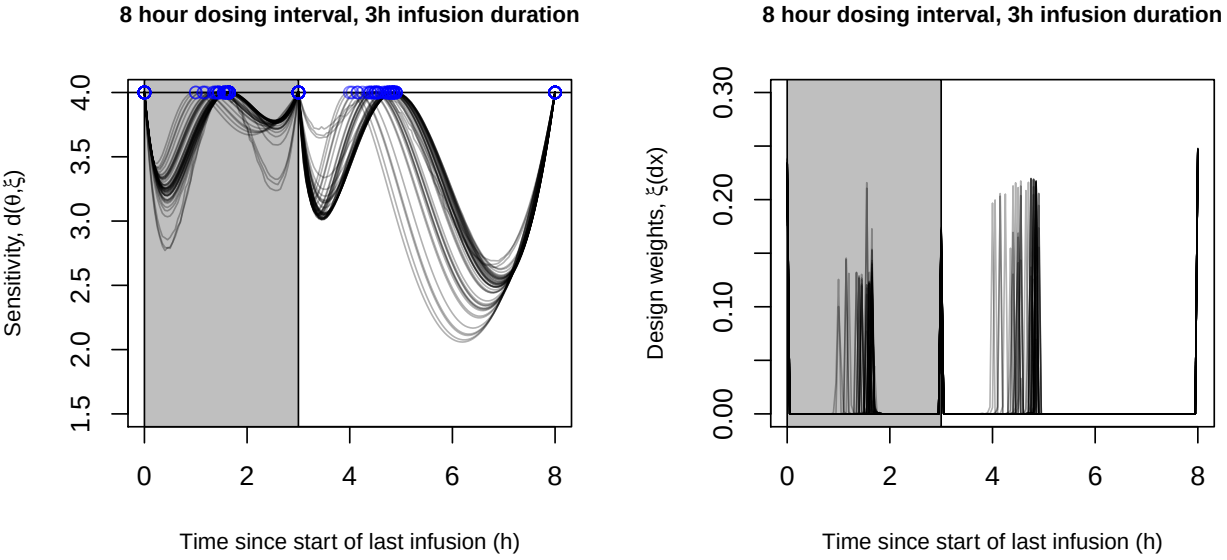


Figure 3: Example concentration curve for PK parameters drawn from prior distribution with simulated data.



Using this model, we conduct what is, in essense, a simulation study in which we identify the optimal design corresponding to a hypothetical population of 50 patients. In theory, each patient's pharmacokinetic profile is fully characterized by a set of parameters drawn from the prior distribution specified above. The response function, $\eta(x, \theta)$, is given by the solution to the differential equation model for the central compartment, which then describes the drug concentration in the bloodstream as a function of time. The gradient function and information matrix are numerically approximated and we initialize the algorithm described in Section 5.2 with a design that assigns equal weight to all time points. For each simulated patient, the algorithm is iterated until convergence and the resulting sensitivity function is checked to verify that its maximum on the support of x is equal to $m = 4$. The equivalence theorems provided earlier provide that this is equivalent to an optimal design (i.e. one

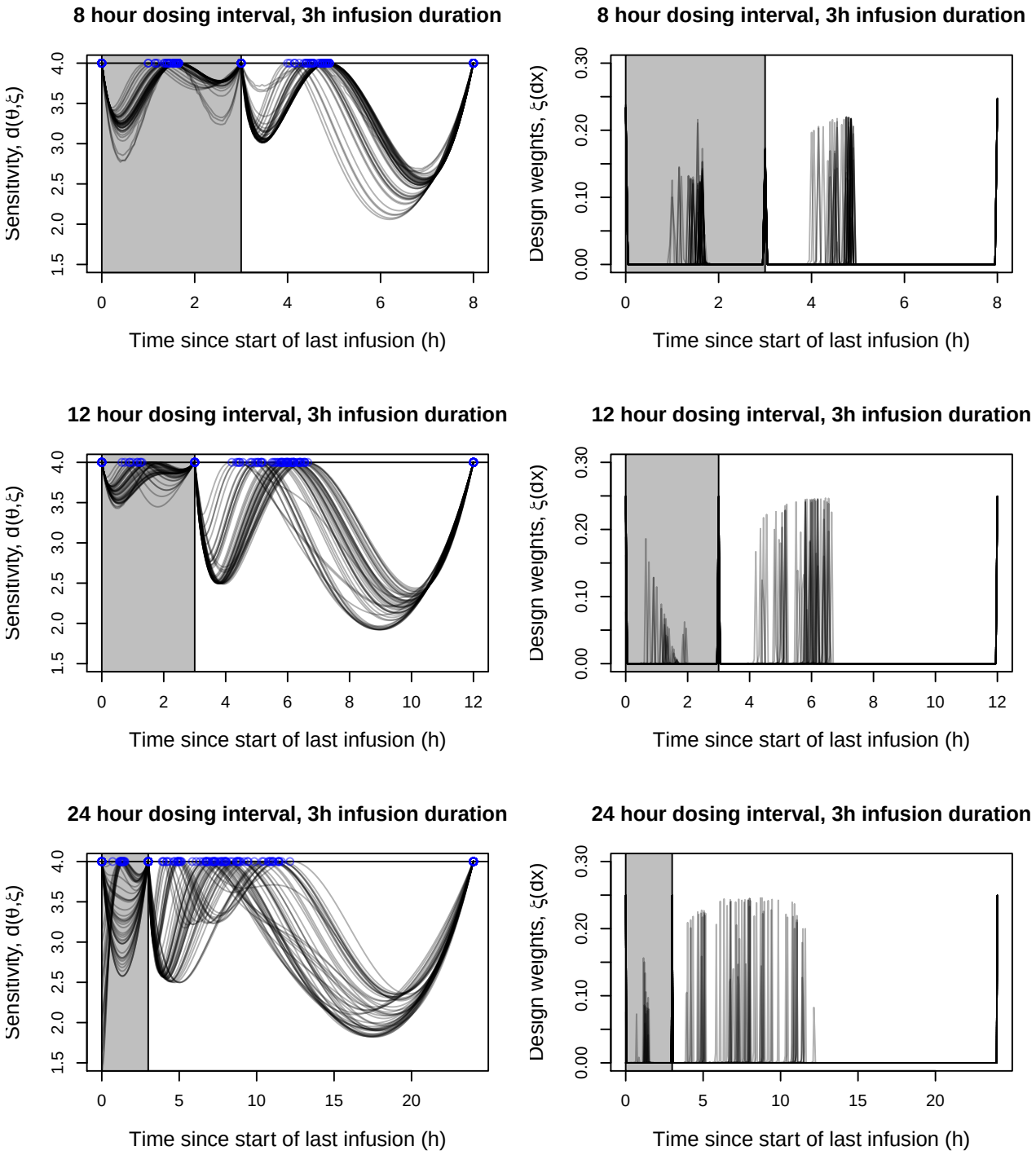


Figure 4: D-optimal design points at different dosing intervals for 5th infusion at 3 hour infusion duration (shaded grey). Each line represents a draw from the prior distribution of PK parameters. Optimal design points are colored blue. Left hand panels show the sensitivity function values associated with the optimal design for each set of parameters. Right hand panels show the corresponding design weights for each optimal design.

that minimizes the determinant of the variance-covariance matrix). The results of these simulations are shown in Figure 4.

The sensitivity functions in the left hand panels of Figure 4 show that for each hypothetical patient there exist 5 design points for the 8 hour and 12 hour dosing intervals, and either 4 or 5 design points in the 24 hour dosing interval. The blue points in Figure 4 indicate the locations at which the sensitivity

function is equal to m. The right hand panels show the design weights associated with the set of 50 optimal designs (one for each set of parameters). In all panels of the figure, lines are drawn with a degree of transparency, such that darker areas indicate agreement across sets of parameters. For example, in the top left panel of Figure 4, design points are located at 0, 3, and 8 hours following the beginning of the final infusion across all 50 "patients." The final two design points are placed roughly at 1.75 and 4.5 hours and show sensitivity to parameter specification.

The portion shaded grey indicates the time during which the drug was administered intravenously. In all dosing scenarios, the same mass of drug was given to each patient, but was given at different rates to account for varing sizes of the dosing windows. In general, the most consistently informative points for each patient occur at the beginning and end of the dosing interval, as well as at the end of the infusion period. The remaining design points occur mid-infusion and after approximately a third of the time between the end of the infusion and the end of the dosing interval has passed, though these design points demonstrate variability between patients.

7 Discussion

7.1 Limitations

There are several limitations of the theory discussed in this paper. The most obvious is that the number of design points is a function of the model under consideration and – in the PK example – the dosing schedule. Consequently, while the theory may provide the optimal weight assigned to each design point, for fixed sample sizes that are not multiples of the number of design points, it is unclear where the remaining measurements should be taken. Box and Lucas (1959) suggest alternately using additional observations to estimate additional parameters in a more general model as a means of model-checking, or to better estimate specific areas of the response surface that the investigator (perhaps due to intuition) believes to be poorly estimated. For a given sample size, however, we can measure the information loss due to discretization of the design. This can be done simply by calculating the relative D-efficiency as follows:

$$D\text{-eff}(\xi) = \frac{|M(\xi,\theta)|^{1/k}}{|M(\xi^*,\theta)|^{1/k}} \tag{27}$$

Using this approach, we can see how various designs with a fixed N compare to the optimal design and choose the one that minimizes the relative information loss if we are confident in our model and parameter values. It is worth emphasizing, however, that there are various reasons that one may not wish to choose the design that minimizes the relative information loss. In addition to the considerations of Box and Lucas, we may wish to emphasize robustness in choice of θ values. The sampling approach used in the PK example above illustrates that not all design points are equally robust to parameter misspecification. Consequently, one may reasonably wish to choose a design that gives higher weight to the support points that were invariant to θ. Approaches such as the minimax and averaging methods do not allow one to view this sensitivity to parameter misspecification.

A additional issue common to nonlinear regression models is that the space of the response function, $\eta(x,\theta)$, is curved with respect to θ, allowing for situations in which $||\eta(x,\theta) - \eta(x,\theta_0)||$ is small even when the distance between θ and θ_0 is large. In practice, this translates to the frequent presence of local optima which can cause the estimator to be non-unique or highly sensitive to small changes in the observations. Pázman and Pronzato (2013) introduce a generalized form of the E-optimality criterion (as well as c-optimality and G-optimality, which are not discussed in this paper) that mitigates this risk by incorporating a Euclidian distance measure between θ and θ_0 into the optimality criterion.

7.2 Extensions

The optimal design theory presented has been extended in diverse ways which we will only briefly mention here. Federov and Leonov (2014) show that constraints, such as the costs associated with the collection of data points, can be incorporated into the existing theory. The general principal behind doing so is to define a cost function associated with each design point

$$\Phi(\xi_N, \theta) = \sum_{i=1}^{n} r_i \phi(x_i, \theta)$$

which is then incorporated into the optimality criterion:

$$\xi^* = \underset{\xi}{\operatorname{argmin}} \Psi \left[\frac{M(\xi, \theta)}{\Phi(\xi, \theta)} \right]$$

Additional applications include the identification design points to simutaneously discriminate between a set of candidate models and estimate the parameters of the selected model. This is achieved through the use of the T-optimality criterion, in which one posits a true model and a set of rival models and chooses the design so as to maximize the sum of squares for lack of fit for the rival model(s) (Atkinson & Fedorov, 1975a,b). One then uses a likelihood ratio test to select a model.

Adaptive designs based on repeated experiments have been developed and are commonly used in dose-finding clinical trials (Federov & Leonov, 2014; Lane, 2013). During adaptive designs, the parameters at each stage are reestimated using data from prior stages to find updated design points. Given that the identification of an optimal design for nonlinear models depends on the model parameters, θ, which are rarely (if ever) known in advance, when repeated experimentation is possible, adaptive designs typically provide a superior alternative to locally optimal designs.

Extensions have also been made into the optimal design of experiments for nonparametric models, in which the form of the response function is unspecified beyond smoothness and regularity properties. Müller (1984) introduced a locally optimal design criterion for nonparametric (kernel density) regression models, which was extended by Biedermann and Dette (2000) to allow for a minimax approach. The latter authors find that in certain cases, uniform weights along $\mathcal{X}$ is optimal. In these scenarios, since kernel density models are parameterized by a bandwidth tuning parameter, a "locally optimal" design refers to one in which the bandwidths are known. In practice, if bandwidths are permitted to vary along the support of the design variable, an optimal design will depend on the selection of the best values for these bandwidths, which are typically unknown.

According to Pronzato & Pázman (2013), since the response function is unknown in nonparametric scenarios, the optimal design is typically that which places design points so as to "have a space-filling property." When our set of design variables is one dimensional and we use a euclidian distance measure, which we may choose to use when we have no prior information regarding the value of the response function over support of our design variable, this translates to sampling uniformly across $\mathcal{X}$. In higher dimensions, however, "space-filling" is not as simple and may be defined in various ways (e.g. which is preferable, a square grid or a triangular grid?) (Pronzato & Müller, 2011). For example, one may choose to maximize the distance (on any measure) between each point and its closest neighbor, in what is called a "maximin-distance design," or one may minimize the maximum distance between all points and their closest neighbor in a "minimax-distance design." Many additional approaches exist and, as before, what can be considered optimal will depend on one's interests.

Many questions of interest in statistics do not allow for an experimental setting in which the researcher is able to collect data at pre-determined design points. For those that do, however, the identification of points that are likely to confer the most information to a researcher's candidate model can result in increased estimate efficiency and is worth consideration in the design stage of a trial or experiment.

References

[1] Atkinson, A. C. & Fedorov, V. V. The Design of Experiments for Discriminating Between two Rival Models. *Biometrika*, Vol. 62, No. 1 (Apr., 1975), pp. 57-70

[2] Atkinson, A. C. & Fedorov, V. V. Optimal Design: Experiments for Discriminating between Several Models. *Biometrika*, Vol. 62, No. 2 (Aug., 1975), pp. 289-303

[3] Biedermann, S. & Dette H. Minimax optimal designs for nonparametric regression - a further optimality property of the uniform distribution. *In A. Atkinson, P. Hackl, and W. Müller (Eds.), mODa'6 – Advances in Model–Oriented Design and Analysis, Proc. 6th Int. Workshop, Puchberg/Schneberg (Austria), Heidelberg, pp. 13–20. Physica Verlag.*

[4] Box, G. E. P. & Lucas, H. L. (1959) Design of Experiments in Non-Linear Situations. *Biometrika.* 46:77-90.

[5] Fedorov, V. V., & Leonov, S. L. (2014). Optimal design for nonlinear response models. Boca Raton: CRC Press/Taylor & Francis Group.

[6] Kiefer, J. & Wolfowitz, J. (1960). The equivalence of two extremum problems. *Canad. J. Math.*, 12:363–366

[7] Lane, A., Yao, P., Flournoy, N. (2013). Information in a two-stage adaptive optimal design *Journal of Statistical Planning and Inference* 144 (2014) 173–187

[8] Müller, Hans-Georg (1984). Optimal Designs for Nonparametric Kernel Regression *Statistics & Probability Letters 2* (1984) 285-290

[9] Pázman A. & Pronzato L. (2013) Extended Optimality Criteria for Optimum Design in Nonlinear Regression. Advances in Model-Oriented Design and Analysis. Contributions to Statistics. Springer, Heidelberg.

[10] Petersen, K. B. & Pedersen, M. S. (2012). The Matrix Cookbook. Technical University of Denmark.

[11] Pronzato L. & Pázman A. (2013) Design of Experiments in Nonlinear Models: Asymptotic Normality, Optimality Criteria and Small-Sample Properties. Springer-Verlag, New York

[12] Pronzato L. & Müller, W. (2012). Design of computer experiments: space filling and beyond. Springer Verlag (Germany), 2012, 22 (3), pp.681-701.

[13] von Gilbert Koch, vorgelegt (2012). Modeling of Pharmacokinetics and Pharmacodynamics with Application to Cancer and Arthritis Department of Mathematics and Statistics, University of Konstanz `http://www.math.uni-konstanz.de/numerik/personen/koch/papers/Diss_Koch_Final.pdf`

8 Appendix

8.1 Derivations of Information Matrices

Given the model

$$y = \eta(x, \theta) + \epsilon$$
$$\epsilon \sim N(0, \sigma^2)$$

the log likelihood function is

$$l(\theta) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{-(y - \eta(x, \theta)^2)}{2\sigma^2}$$

and corresponding score function

$$\frac{\partial l\theta}{\partial \theta} = S(\theta) = \frac{(y - \eta(x, \theta))\nabla\eta}{\sigma^2}$$

In the linear case, $\nabla\eta = x$, and the score function reduces to

$$S(\theta) = \frac{(y - x^T\theta)x}{\sigma^2}$$

The Hessian is given by

$$\frac{\partial S(\theta)}{\partial \theta} = H(\theta) = -\frac{xx^T}{\sigma^2}$$

And the Fisher Information Matrix by the negative expectation of the Hessian

$$I(\theta) = -E_\theta[H(\theta)] = \frac{xx^T}{\sigma^2}$$

Which coincides with the information matrix for the least-squares estimator. In the non-linear case, we make a first-order Taylor series approximation about the true value of θ, θ_t:

$$\eta(x, \theta) \approx \eta(x, \theta_t) + (\theta - \theta_t)^T\nabla\eta$$

Substituting this approximation into the initial model

$$\begin{aligned} y - \epsilon &= \eta(x, \theta_t) + (\theta - \theta_t)^T\nabla\eta \\ y - \eta(x, \theta_t) + \theta_t^T\nabla\eta &= \theta^T\nabla\eta + \epsilon \\ y^* &= \theta^T\nabla\eta + \epsilon \end{aligned}$$

Which is linear in θ and of the same form as the linear model described above. The steps to calculate the Fisher information matrix are identical and we end up with

$$I(\theta) \approx \frac{(\nabla\eta)^T(\nabla\eta)}{\sigma^2}$$

Similarly, for nonlinear least-squares

$$\hat{\theta} = (\nabla\eta^T\nabla\eta)^{-1}\nabla\eta^T y^*$$

$$V(\hat{\theta}) \approx \sigma^2(\nabla\eta^T\nabla\eta)^{-1}$$

In both the maximum likelihood and the least-squares approaches, iterative algorithms can be used to solve for $\hat{\theta}$ and $V(\hat{\theta})$, where the estimate at each step is substitued in for the "true" value, θ_t. The accuracy of the estimated information matrix in the non-linear case is obviously dependent upon the appropriateness of the first order Taylor series approximation.

8.2 Derivation of sensitivity function for D-criterion

As stated previously, the directional derivative associated with the fourth necessary and sufficient condition for optimality takes the form

$$\int_{\mathcal{X}} \psi(x,\xi)\bar{\xi}(dx)$$

The general form of the sensitivity function is

$$\varphi(x,\xi) = -\psi(x,\xi) + C$$

where C is equal to the number of parameters estimated, m, for the D-criterion and is equal to $\Psi(\xi)$ for all other criterion. An α-weighted design with respect to arbitrary designs $\xi, \bar{\xi}$ is given as

$$\xi_\alpha = (1-\alpha)\xi + \alpha\bar{\xi} = \xi + \alpha(\bar{\xi} - \xi)$$

with a corresponding information matrix

$$M_\alpha = (1-\alpha)M(\xi) + \alpha M(\bar{\xi}) = M(\xi) + \alpha(M(\bar{\xi}) - M(\xi))$$

To find $\psi(x,\xi)$ for a specific optimality criterion, we calculate the derivative of our optimality criterion for an α-weighted design with respect to α and evaluate the limit as $\alpha \to 0$.

$$
\begin{aligned}
\Psi_\alpha &= -ln|M_\alpha| \\
\frac{d\Psi_\alpha}{d\alpha} = \Psi'_\alpha &= -tr\left\{M_\alpha^{-1}\frac{dM_\alpha}{d\alpha}\right\} \quad \text{(Matrix cookbook, eq.43)} \\
&= -tr\left\{M_\alpha^{-1}[M(\bar{\xi}) - M(\xi)]\right\} \\
&= -tr\left\{[M(\xi) + \alpha(M(\bar{\xi}) - M(\xi)]^{-1}[M(\bar{\xi}) - M(\xi)]\right\} \\
\lim_{\alpha\to 0}\Psi'_\alpha = \dot{\Psi}_\alpha &= -tr\left\{M(\xi)^{-1}[M(\bar{\xi}) - M(\xi)]\right\} \quad \text{(Interchange limit and summation)} \\
&= -tr\left\{M(\xi)^{-1}\left[\int \mu(x)\bar{\xi}(dx) - \int \mu(x)\xi(dx)\right]\right\} \\
&= -tr\left\{M(\xi)^{-1}\left[\int \mu(x)\bar{\xi}(dx) - \iint \mu(x)\bar{\xi}(dx)\xi(dx)\right]\right\} \quad \left(\int \bar{\xi}(dx) = 1\right) \\
&= -tr\left\{M(\xi)^{-1}\left[\int \bar{\xi}(dx)\left[\mu(x) - \int \mu(x)\xi(dx)\right]\right]\right\} \\
&= -tr\left\{\int \bar{\xi}(dx)\left[M(\xi)^{-1}\mu(x) - M(\xi)^{-1}M(\xi)\right]\right\} \\
&= \int \bar{\xi}(dx)\left[-tr(M(\xi)^{-1}\mu(x)) + m\right] \\
&= \int \bar{\xi}(dx)\left[m - tr(M(\xi)^{-1}\mu(x))\right]
\end{aligned}
$$

Therefore,

$$\psi(x,\xi) = m - tr(M(\xi)^{-1}\mu(x))$$

Giving us the sensitivity function for the D-criterion:

$$\varphi(x,\xi) = tr(M(\xi)^{-1}\mu(x))$$

The limit and trace operators can be interchanged during the derivation by the Dominated Convergence Theorem. The trace of $f_\alpha = M_\alpha^{-1}[M(\bar{\xi}) - M(\xi)]$ is a summation of m positive and finite values, all of which can be bounded by merely adding a constant to each: $g_\alpha = f_\alpha + c$. Since f_α can be bounded by another function, g_α, with a finite sum, the theorem applies and we have

$$\lim_{\alpha\to 0} tr(f_\alpha) = tr(f)$$

8.3 Equivalence of sensitivity function and variance of predicted response

We wish to show that the sensitivity function for the D-criterion, given as

$$tr[M^{-1}(\eta,\theta)\mu(x,\theta)]$$

is approximately equivalent to the variance of the predicted response, $\eta(x,\hat{\theta})$. We begin by evaluating the variance of the Taylor series approximation to the predicted response (i.e. using the multivariate delta method).

$$
\begin{aligned}
\eta(x,\hat{\theta}) &\approx \eta(x,\theta) + (\hat{\theta}-\theta)\nabla\eta \\
V(\eta(x,\hat{\theta})) &\approx V(\eta(x,\theta) + (\hat{\theta}-\theta)\nabla\eta) \\
&= V(\hat{\theta}\nabla\eta) \\
&= \nabla\eta^T D \nabla\eta \\
&= (\nabla\eta_1, ..., \nabla\eta_k) \begin{pmatrix} D_{11} & \dots & D_{1k} \\ \vdots & \ddots & \vdots \\ D_{k1} & \dots & D_{kk} \end{pmatrix} \begin{pmatrix} \nabla\eta_1 \\ \vdots \\ \nabla\eta_k \end{pmatrix} \\
&= \nabla\eta_1 \sum_{j=1}^{k} \nabla\eta_j D_{j1} + ... + \nabla\eta_k \sum_{j=1}^{k} \nabla\eta_j D_{jk} \\
&= \sum_{i=1}^{k} D_{ii}\nabla\eta_i^2 + \sum_i \sum_{j\neq i} D_{ij}\nabla\eta_i\nabla\eta_j
\end{aligned}
$$

We can then see that this expression is identical to the sensitivity function for the D-criterion

$$
\begin{aligned}
tr[M^{-1}(\eta,\theta)\mu(x,\theta)] &= tr[M^{-1}(\eta,\theta)\nabla\eta\nabla\eta^T] \\
&= tr\left(\begin{pmatrix} D_{11} & \dots & D_{1k} \\ \vdots & \ddots & \vdots \\ D_{k1} & \dots & D_{kk} \end{pmatrix} \begin{pmatrix} \nabla\eta_1^2 & \dots & \nabla\eta_1\nabla\eta_k \\ \vdots & \ddots & \vdots \\ \nabla\eta_k\nabla\eta_1 & \dots & \nabla\eta_k^2 \end{pmatrix} \right) \\
&= tr(\mathbf{S})
\end{aligned}
$$

The diagonal elements of the matrix $\mathbf{S}$ are given

$$
\begin{aligned}
\mathbf{S}_{11} &= D_{11}\nabla\eta_1^2 + D_{12}\nabla\eta_2\nabla\eta_1 + ... + D_{1k}\nabla\eta_k\nabla\eta_1 \\
\mathbf{S}_{22} &= D_{21}\nabla\eta_1\nabla\eta_2 + D_{22}\nabla\eta_2^2 + ... + D_{1k}\nabla\eta_k\nabla\eta_2 \\
&\vdots \\
\mathbf{S}_{kk} &= D_{k1}\nabla\eta_1\nabla\eta_k + ... + D_{k,k-1}\nabla\eta_{k-1}\nabla\eta_k + D_{kk}\nabla\eta_k^2
\end{aligned}
$$

When summed, these equal

$$
tr(\mathbf{S}) = \mathbf{S}_{11} + \mathbf{S}_{22} + ... + \mathbf{S}_{kk} = \sum_{i=1}^{k} D_{ii}\nabla\eta_i^2 + \sum_i \sum_{j\neq i} D_{ij}\nabla\eta_i\nabla\eta_j
$$

8.4 Solution to differential equation model

Here we derive the solution to the two-compartmental differential equation model presented for a single dosing period. The general approach to the derivation follows the example of vorgelegt von Gilbert Koch (2012), who uses the Laplace transform method to find the solution for the homogenous linear system (i.e. without the drug infusion rate, k_R). The concentration across multiple dosing periods is given by a piecewise function with the initial concentration in the primary compartment given by the concentration at the conclusion of the prior dosing period.

The system of equations, expressed in terms of volume rather than concentration and without the initial infusion is given

$$\begin{aligned}\frac{dx_1}{dt} &= x_1'(t) = -(k_{10}+k_{12})x_1(t) + k_{21}x_2(t) + k_R \\ \frac{dx_2}{dt} &= x_2'(t) = k_{12}x_1(t) - k_{21}x_2(t)\end{aligned} \tag{28}$$

and can be expressed in matrix form as

$$x'(t) = Ax(t) + R$$

In which

$$x'(t) = \begin{pmatrix} x_1'(t) \\ x_2'(t) \end{pmatrix} \quad A = \begin{pmatrix} -(k_{10}+k_{12}) & k_{21} \\ k_{12} & -k_{21} \end{pmatrix} \quad x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} \quad R = \begin{pmatrix} k_R \\ 0 \end{pmatrix}$$

The Laplace transform of a function, $f(t)$ is given

$$\mathcal{L}(f(t)) = \int_0^\infty f(t)e^{-st}dt \equiv F(s)$$

We note the following properties to be used:

1. $\mathcal{L}(A) = \frac{A}{s}$ by direct calculation.
2. $\mathcal{L}(Ax(t)) = AX(s)$ by definition.
3. $\mathcal{L}(x'(t)) = sX(s) - x(0)$ by direct calculation using integration by parts.

We now apply the Laplace transform to both sides

$$\begin{aligned} x'(t) &= Ax(t) + R \\ \mathcal{L}(x'(t)) &= \mathcal{L}(Ax(t)) + \mathcal{L}(R) \\ sX(s) - x(0) &= AX(s) + R(s) \quad R(s)^T = (k_R/s, 0) \\ (sI - A)X(s) &= x(0) + R(s) \\ X(s) &= (sI - A)^{-1}(x(0) + R(s)) \end{aligned}$$

We define

$$sI - A = M(s) = \begin{pmatrix} s + k_{10} + k_{12} & -k_{21} \\ -k_{12} & s + k_{21} \end{pmatrix}$$

such that

$$M^{-1}(s) = \frac{1}{(s+k_{10}+k_{12})(s+k_{21}) - k_{12}k_{21}} \begin{pmatrix} s + k_{21} & k_{21} \\ k_{12} & s + k_{10} + k_{12} \end{pmatrix}$$

We note that

$$\begin{aligned}(s+k_{10}+k_{12})(s+k_{21}) - k_{12}k_{21} &= s^2 + s(k_{10}+k_{12}+k_{21}) + k_{10}k_{21} \\ &= (s+\alpha)(s+\beta)\end{aligned}$$

where by the quadratic formula,

$$\alpha, \beta = \frac{1}{2}\left[k_{10} + k_{12} + k_{21} \pm \sqrt{(k_{10} + k_{12} + k_{21})^2 - 4k_{10}k_{21}}\right]$$

and additionally that $-\alpha, -\beta$ are the eigenvalues of A, λ_1, λ_2.

$$\begin{aligned} M^{-1}(s)(x(0) + R(s)) &= \frac{1}{(s+\alpha)(s+\beta)} \begin{pmatrix} s + k_{21} & k_{21} \\ k_{12} & s + k_{10} + k_{12} \end{pmatrix} \begin{pmatrix} k_R/s + x_1(0) \\ x_2(0) \end{pmatrix} \\ &= \frac{1}{(s+\alpha)(s+\beta)} \begin{pmatrix} (s + k_{21})(k_R/s + x_1(0)) + k_{21}x_2(0) \\ k_{12}(k_R/s + x_1(0)) + (s + k_{10} + k_{12})x_2(0) \end{pmatrix} \end{aligned}$$

To take the inverse Laplace transform of $M^{-1}(s)(x(0)+R(s))$, we can use Heaviside's Theorem, which states

$$\mathcal{L}^{-1}\left(\frac{p(s)}{q(s)}\right) = \sum_{i=1}^{n} \frac{p(\lambda_i)}{q'(\lambda_i)} e^{\lambda_i t}$$

where λ_i is the ith eigenvalue of A. To apply this theorem, we note that $q(s) = (s+\alpha)(s+\beta)$ and $q'(s) = 2s + \alpha + \beta$.

we then can solve for the mass in the first compartment

$$\begin{aligned} x_1(t) = \mathcal{L}^{-1}(X_1(s)) &= \mathcal{L}^{-1}(M^{-1}(s)(x(0) + R(s))) \\ &= \left[\frac{sx_1(0) + \frac{k_R k_{21}}{s}}{q'(s)}\Big|_{s=-\alpha} e^{-\alpha t} + \frac{sx_1(0) + \frac{k_R k_{21}}{s}}{q'(s)}\Big|_{s=-\beta} e^{-\beta t}\right] + \\ &\quad (k_R + k_{21}(x_1(0) + x_2(0)))\left[q'(s)^{-1}\Big|_{s=-\alpha} e^{-\alpha t} + q'(s)^{-1}\Big|_{s=-\beta} e^{-\beta t}\right] \\ &= \left[\frac{-\alpha x_1(0) + \frac{k_R k_{21}}{-\alpha}}{\beta - \alpha} e^{-\alpha t} + \frac{-\beta x_1(0) + \frac{k_R k_{21}}{-\beta}}{\alpha - \beta} e^{-\beta t}\right] + \\ &\quad \left[\frac{k_R + k_{21}(x_1(0) + x_2(0))}{\beta - \alpha} e^{-\alpha t} + \frac{k_R + k_{21}(x_1(0) + x_2(0))}{\alpha - \beta} e^{-\beta t}\right] \\ &= \frac{\alpha k_R - k_R k_{21} + \alpha[k_{21}(x_1(0) + x_2(0)) - \alpha x_1(0)]}{\alpha(\beta - \alpha)} e^{-\alpha t} + \\ &\quad \frac{\beta k_R - k_R k_{21} + \beta[k_{21}(x_1(0) + x_2(0)) - \beta x_1(0)]}{\beta(\alpha - \beta)} e^{-\beta t} \end{aligned}$$

At the start of the first infusion the inital mass in each compartment is equal to 0, such that $x_1(0) = x_2(0) = 0$ and the solution simplifies to

$$x_1(t) = \frac{k_R(k_{21} - \alpha)}{\alpha(\alpha - \beta)} e^{-\alpha t} + \frac{k_R(k_{21} - \beta)}{\beta(\beta - \alpha)} e^{-\beta t}$$

The solution can be expressed in terms of the concentration within the central compartment at time t merely by dividing the mass of the drug in the central compartment at time t, $x_1(t)$, by the volume of the central compartment, v_1:

$$c_1(t) = \frac{x_1(t)}{v_1}$$

9 Additional Graphics