A comparison of sequential design procedures for discriminating enzyme kinetic models

Elham Yousefi *
Department of Applied Statistics, Johannes Kepler University Linz, Austria

Abstract

Optimal designs for model discrimination often depend on the true model and its parameters. To proceed sequentially is one of the alternatives to tackle this dependence. In this study we consider different sequential design strategies based on T-, D_s- and δ-optimality for discriminating between different enzyme kinetic profiles and make a comparison between their performances including one of the speed of their convergences.

Keywords: Nonlinear Regression, Sequential designs, Discriminating designs, Convergence rate.

1 Introduction

Many optimal design procedures for discrimination (including T-optimal designs) depend on which model is the true one and also on the parameters of that model and therefore are locally optimum. One option to tackle this dependence is to apply a sequential procedure. Two rival models are considered and it is often assumed that one of the presented models represents the true physical mechanism under which the real data are generated (i.e., the data generating process). Note that the situation becomes slightly more complex when we assume that neither of the two models might be the true one. In the former sense the rival models are

$$\begin{aligned} y_i &= \eta_0(\boldsymbol{\theta}_0, \mathbf{x}_i) + \epsilon_i, \qquad i = 1, \ldots, N, \quad \text{and} \\ y_i &= \eta_1(\boldsymbol{\theta}_1, \mathbf{x}_i) + \epsilon_i, \qquad i = 1, \ldots, N, \end{aligned}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^T$ is the vector of m unknown parameters, $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}_+^m$, $\boldsymbol{\Theta}$ is a compact set of admissible parameter values and $\mathbf{x}$ denotes the design variable(s). Further y denotes the observation, $\eta(\boldsymbol{\theta}, \mathbf{x})$ is the expected response, a nonlinear function of the unknown parameters and the design variables and the random errors are i.i.d. with $\mathcal{N}(0, \sigma^2)$.

2 Methodology

In order to conduct a sequential experiment, let N_0 observations $y_0, \ldots, y_{N_0}$ of the response be given. The goal is to find the next setting for the design variables (next trial) which best

*e-mail: elham.yousefi@jku.at

discriminates between two models. Utilizing the one which maximizes the residual sum of squares of the incorrect model using the nonlinear least squares estimates from the runs completed until then, yields the classic design strategy proposed in Hunter and Reiner (1965) and Atkinson and Fedorov (1975).

1. Let an initial exact nonsingular design ξ_{N_0} with N_0 observations $y_i, i = 1, \ldots, N_0$ be given. From these observations find the nonlinear least square parameter estimates of the models $(\hat{\boldsymbol{\theta}}_{0N_0}, \hat{\boldsymbol{\theta}}_{1N_0})$

$$\sum_{i=1}^{N_0} (y_i - \widehat{y_{ji}})^2 = \inf_{\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j} \sum_{i=1}^{N_0} (y_i - \eta_j(\boldsymbol{\theta}_j, \mathbf{x}_i))^2 \quad (\widehat{y_{ji}} = \eta_j(\hat{\boldsymbol{\theta}}_{jN_0}, \mathbf{x}_i), \quad j = 0, 1).$$

2. The next point $\mathbf{x}_{N_0+1}$ is chosen as $\mathbf{x}_{N_0+1} = \arg\max_{\mathbf{x} \in \mathfrak{X}} \left\{ \eta_0(\hat{\boldsymbol{\theta}}_{0N_0}, \mathbf{x}) - \eta_1(\hat{\boldsymbol{\theta}}_{1N_0}, \mathbf{x}) \right\}^2$.

3. The $(N_0 + 1)$th observation is taken at $\mathbf{x}_{N_0+1}$.

4. Steps 1 to 3 are repeated.

This sequential procedure will lead to an asymptotically T-optimal design. Such procedures are of high importance when there is a predefined budget to perform experiments in application phases. Sequential D_s-optimal designs could similarly be used for model discrimination, with the difference that the first and the second stages in the above sequential strategy are changed to fit the D_s-optimality criterion (Atkinson and Cox, 1974; Atkinson et al., 2007).

The last discrimination procedure investigated sequentially here is δ-optimality (Harman and Müller (2020)), which is inherently a symmetric criterion based on linearizing the rival models while considering flexible nominal parameter sets. We now need to alter just the second stage of the above sequential strategy, while the other stages remains the same, as follows.

$$(y_i)_{i=1}^{N_0+1} \approx \mathbf{F}_u(\mathcal{D})\boldsymbol{\theta}_u + \mathbf{a}_u(\mathcal{D}) + \epsilon, \quad \mathbf{a}_u(\mathcal{D}) = (\eta_u(\hat{\boldsymbol{\theta}}_u, \mathbf{x}_i))_{i=1}^{N_0+1} - \mathbf{F}_u(\mathcal{D})\hat{\boldsymbol{\theta}}_u, \quad u = 0, 1,$$

where $\mathcal{D} = (\mathbf{x}_1, \ldots, \mathbf{x}_{N_0+1})$ is an exact design of size $N_0 + 1$, $\mathbf{F}_u(\mathcal{D})$ is the $(N_0 + 1) \times m$ matrix of partial derivatives at $\hat{\boldsymbol{\theta}}_u$, the respective nominal values (parameter estimates). According to above notations, the linearized distance criterion is (see Harman and Müller (2020) for more details)

$$\delta_r(\mathcal{D}) = \inf_{\boldsymbol{\theta}_0 \in \hat{\boldsymbol{\Theta}}_0^{(r)}, \boldsymbol{\theta}_1 \in \hat{\boldsymbol{\Theta}}_1^{(r)}} \delta(\mathcal{D} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1). \tag{1}$$

$$\delta(\mathcal{D} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = \| \mathbf{a}_0(\mathcal{D}) + \mathbf{F}_0(\mathcal{D})\boldsymbol{\theta}_0 - \{\mathbf{a}_1(\mathcal{D}) + \mathbf{F}_1(\mathcal{D})\boldsymbol{\theta}_1\} \|,$$

where $\hat{\boldsymbol{\Theta}}_0^{(r)} \subseteq \mathbb{R}^m$, $\hat{\boldsymbol{\Theta}}_1^{(r)} \subseteq \mathbb{R}^m$ are called the flexible nominal sets. For a set $\mathfrak{D}$ of all (N_0+1)-point designs, a design $\mathcal{D}^* \in \mathfrak{D}$ will be called δ-optimal, if

$$\mathcal{D}^* \in \arg\max_{\mathcal{D} \in \mathfrak{D}} \delta_r(\mathcal{D}). \tag{2}$$

Therefore Eq. (1) is computed for all designs of the size $N_0 + 1$ (the design of size N_0 so far plus one candidate point, together of size $N_0 + 1$). This is done for all points on the grid and the design of size $N_0 + 1$ called $\mathcal{D}^*$ is chosen which maximizes the δ-criterion as in 2.

3 Result

The above sequential methods are now applied to competitive and non-competitive inhibition models, which are widely used in drug discovery (Copeland, 2005) and also investigated by many authors in optimal design (Bogacka et al., 2011; Atkinson, 2012; Harman and Müller, 2020) and are respectively defined as

$$y = \frac{\theta_V x_S}{\theta_M \left(1 + \frac{x_I}{\theta_K}\right) + x_S} + \epsilon, \tag{3}$$

$$y = \frac{\theta_V x_S}{(\theta_M + x_S) \left(1 + \frac{x_I}{\theta_K}\right)} + \epsilon, \tag{4}$$

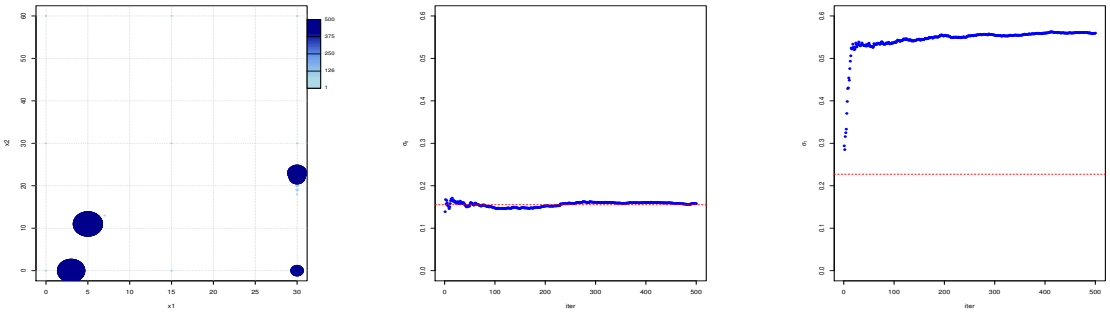
where $\mathbf{x} = (x_S, x_I)^T$ denotes the pair of design variables. Atkinson (2011) suggested to combine the above models to form an encompassing model as

$$y = \frac{\theta_V x_S}{\theta_M \left(1 + \frac{x_I}{\theta_K}\right) + x_S \left(1 + \frac{(1-\lambda)x_I}{\theta_K}\right)} + \epsilon, \tag{5}$$

where the fourth parameter is $0 \leq \lambda \leq 1$. $\lambda = 1$ corresponds to the competitive (Eq.(3)) and $\lambda = 0$ to the non-competitive model (Eq.(4)) . $\mathfrak{X} = [0, 30] \times [0, 60]$ is the design space (grid of all candidate points) used here and the search is over a grid of 31×61 pair values of $\mathbf{x}$.

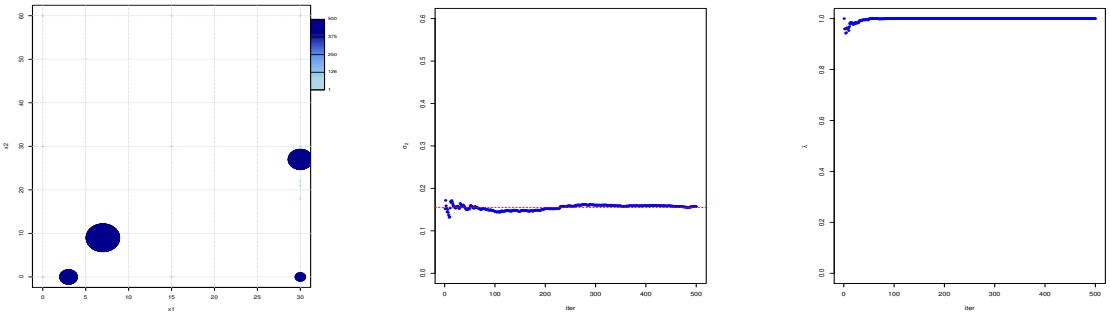
The following plots show the results from the above sequential strategy, where the resulting designs drawn as dark blue circles in Figure 1a are convergent cases for the T-optimal designs when the competitive model is the data generator (with $B = 500$ number of iterations). As it is observed from Figure 1a, it takes some iterations until the sequential designs start replicating over those four points in dark blue. To observe how the approximate optimal discriminating designs for these models and also their log transformed forms look like, one may refer to Yousefi and Müller (2021). Some discrepancies between the optimal designs (compared to the ones in Atkinson (2012)) occur due to differences in initial estimates and the design space. Furthermore, areas of the circles are proportional to their approximate weights, proportion of replications at each support point $\mathbf{x}$. Figure 1b presents the residual standard error estimates of the competitive model, which are completely compatible with their assumed null value $\hat{\sigma} = 0.1553$, drawn as the red dashed horizontal line in the figure. The last Figure 1c shows the residual standard error estimates of the noncompetitive model (basically the model not generating the observations), which is deviating from its null value $\hat{\sigma} = 0.2272$, the red dashed horizontal line. Note that these results confirm that the residual standard error estimates of the true model are unaffected by the choice of design variable levels (Hunter and Reiner (1965)).

The following plots show the results from the D_s sequential procedure, where the interest is in estimation of the discrimination parameter λ in the encompassing model (Eq. (5)). From Figure 2a one can observe a reasonably high convergence rate of the D_s sequential method to get toward the approximate D_s-optimal designs. This suggests that D_s is a fast sequential procedure. Estimated residual standard errors of the encompassing model converge to their null value $\hat{\sigma} = 0.1553$ after a few iterations, as may be observed from Figure 2b. Further, the estimated λ in Figure 2c is mostly equal to one, the value for the competitive model.



(a) (b) (c)

Figure 1: T-sequential procedure, the competitive model is the data generator. (a): sequentially constructed designs, (b) and (c): residual standard error estimates of the competitive and noncompetitive models, respectively.

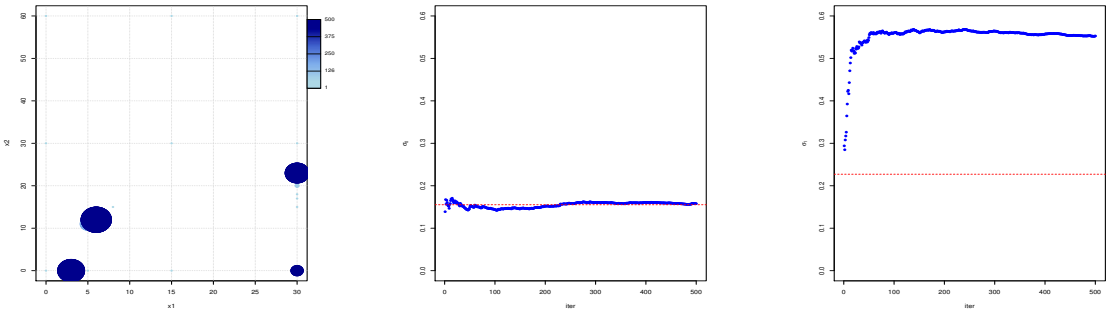


(a) (b) (c)

Figure 2: D_s-sequential procedure, the competitive model is used to generate the obs. (a): sequentially constructed designs, (b): residual standard error estimates of the encompassing model, (c): estimates of λ.

The last sequential procedure investigated here is δ-optimality, where the rival models are considered in the linearized form and the nominal intervals are specifically chosen as $\hat{\mathbf{\Theta}}^{(r)} = [\hat{\theta}_{u1} \pm r\hat{\sigma}_{u1}] \times [\hat{\theta}_{u2} \pm r\hat{\sigma}_{u2}] \times [\hat{\theta}_{u3} \pm r\hat{\sigma}_{u3}]_{u=0,1}$. In the computations we use $r = 1$, but the specific value is not as essential in the static case since we have automatic adaption of the interval lengths by variance estimation. As it may be observed from Figure 3a, this method is also converging reasonably. The interpretations for the estimated residual standard errors of both models in Figures 3b and 3c are similar to those for the other procedures. Note that δ-optimality was originally developed to compute exact and not approximate designs. Nevertheless as can be seen these newly computed sequential designs for δ-optimality provide reasonable limit designs.

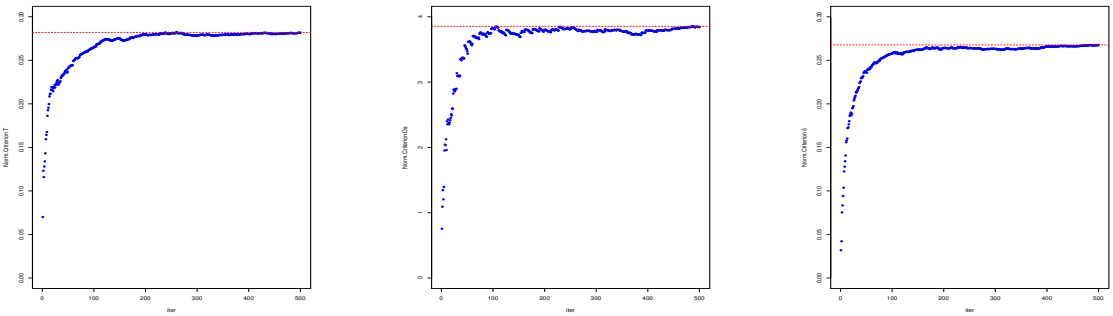
The following figure , Fig. 4 shows the criterion values normalized by the number of observations at each iteration of the sequential procedure (normalized criterion values) which can help to understand how the gathered information, reflected in the criterion values, evolve per observation at each step of the sequential procedure. As it may be observed from this figure for all the three methods the normalized criterion values start settling down (getting smooth) and getting nearer and nearer to their own maximum (drawn as red dashed horizontal lines) from some point on and this suggests that one could stop the sequential procedure sooner than $B = 500$ number of iterations, to save budget, in applications.



(a) (b) (c)

Figure 3: δ-sequential procedure, the competitive model is used to generate the obs. (a): sequentially constructed designs, (b) and (c): residual standard error estimates of the competitive and noncompetitive models, respectively.

The way that the normalized criterion values evolve can help to understand the speed of convergence of each method in a more detailed way by comparing the first time when (the iteration in which) the normalized criterion values level off. This is represented in Table 1. The left block of Table 1 suggests that D_s sequential procedure is slightly faster than the others since it requires less minimum number of observations to reach different quantiles of its maximum (to get nearer and nearer to the maximum information). Similar observations about the other δ and T sequential procedures suggest their reasonable convergence rates. Similar information is provided in the right block of the table, if one assumes to have the noncompetitive model as the data generator. Generally it is observed that all methods perform well.



(a) (b) (c)

Figure 4: Normalized criterion values for all three methods (a): T, (b): D_s, (c): δ

4 Discussion and Conclusion

In this work we considered three sequential design procedures to observe how they behave and in order to compare them. It was seen that all T, D_s and δ-sequential procedures behave reasonable and that specially the D_s-sequential procedure has a relatively higher rate of convergence, compared to T and δ-sequential procedures, by comparing the normalized criterion values. Moreover, the simplicity of computations makes the D_s-sequential procedure even more attractive. Note

Table 1: Minimum number of required observations (Min obs.) to reach different quantiles of the respective maximum for all three methods, left under competitive and right under noncompetitive model.

	Min obs. (competitive)				Min obs. (noncompetitive)			
	50%	75%	90%	95%	50%	75%	90%	95%
T	6	16	73	105	5	16	77	97
D_s	6	28	45	61	7	16	42	74
δ	9	26	58	87	22	60	105	138

that similar results (with slight changes) hold if the noncompetitive model is the data generator. These results can help an experimenter with a fixed and limited budget decide on how to proceed with sequential discriminating designs for enzyme kinetic models.

Acknowledgments I warmly thank my supervisor Werner G. Müller for his constructive comments and discussions throughout the paper.

Supporting information The author's research was fully supported by project grants LIT-2017-4-SEE-001 funded by the Upper Austrian Government, and Austrian Science Fund (FWF): I 3903-N32.

References

Atkinson, A. and V. Fedorov (1975). The design of experiments for discriminating between two rival models. *Biometrika 62*(1), 57–70.

Atkinson, A. C. (2011). Some optimum designs for building enzyme inhibition kinetic models. *Journal of Statistical Theory and Practice 10*, 143–162.

Atkinson, A. C. (2012). Optimum experimental designs for choosing between competitive and non competitive models of enzyme inhibition. *Communications in Statistics-Theory and Methods 41*(13-14), 2283–2296.

Atkinson, A. C. and D. R. Cox (1974). Planning experiments for discriminating between models. *Journal of the Royal Statistical Society: Series B (Methodological) 36*(3), 321–334.

Atkinson, A. C., A. Donev, and R. Tobias (2007). *Optimum Experimental Designs, with SAS (Oxford Statistical Science Series)*. Oxford University Press.

Bogacka, B., M. Patan, P. J. Johnson, K. Youdim, and A. C. Atkinson (2011). Optimum design of experiments for enzyme inhibition kinetic models. *Journal of Biopharmaceutical Statistics 21*(3), 555–572.

Copeland, R. A. (2005). Evaluation of enzyme inhibitors in drug discovery. *Wiley-Interscience*, 34–37.

Harman, R. and W. G. Müller (2020). A design criterion for symmetric model discrimination based on flexible nominal sets. *Biometrical Journal 62*(4), 1090–1104.

Hunter, W. G. and A. M. Reiner (1965). Designs for discriminating between two rival models. *Technometrics 7*(3), 307–323.

Yousefi, E. and W. G. Müller (2021). Impact of the error structure on the design and analysis of enzyme kinetic models. *arXiv preprint arXiv:2103.09563*.