Sequential Design of Experiments with Unknown Covariates

Harvineet Singh hs3673@nyu.edu
New York University

Rumi Chunara rumi.chunara@nyu.edu
New York University

Abstract

We study the problem of optimal design of experiments when the design points are not known apriori or are not under experimenter's control. Such settings occur naturally in longitudinal experiments where the covariates, to be observed in the future, are not available to compute the optimal design beforehand. We discuss the problem in the context of designs for parameter estimation in mixed effects models, which are commonly used for longitudinal data analyses. We propose an approach that predicts the unknown covariates via an autoregressive model and sequentially decides the next design points based on the predicted design criterion. Simulations demonstrate efficiency gains of the proposed approach as compared to choosing design points uniformly at random.

Keywords: Optimal Experiment Design, Longitudinal Data, Mixed Effects Model

1. Introduction

Optimal design of experiments has an extensive literature (Fedorov, 2010; Chaloner and Verdinelli, 1995) which continues to expand with increase in complexity of data (Romero et al., 2013) and experiment goals (Kandasamy et al., 2019). In the standard setup, the experimenter starts by specifying a design space (e.g. a set of protein structures in a protein engineering problem) and a design criterion that quantifies the utility of an experiment (e.g. thermostability of the structure). The goal is to choose design points from the space and conduct corresponding experiments to achieve the optimal value of the design criterion. Importantly, the setup assumes that the design space is fully-observed and the experimenter can choose what design points to experiment on. These assumptions are particularly restrictive in experiments where repeated measurements are taken in time, as discussed next.

Longitudinal studies involve collecting a series of measurements for an experiment unit (say, a human subject) at multiple time points. The study goal usually is to efficiently estimate association between a measured covariate and the response. In such studies, the experimenter can only control some aspects of the design such as what variables to measure and when to take the measurements, but, has no control on the actual values of the measurements. Thus, both covariates and responses are apriori unknown to the experimenter in contrast to the standard setup where only the responses are unknown. Since the design criterion depends on the actual measurements, solving the optimal design problem is challenging in such studies.

Motivating Example Consider a study of sedentary behavior in adults (Kendzor et al., 2016). The study involves repeated measurements of sedentary behavior, physical activity, smoking behavior, perceived stress and other psycho-social variables through self-reports. Suppose, the scientific question of interest is whether perceived stress is related to time spent doing physical activity. That is, we want to quantify the association between two time-varying variables – stress and active time. A characteristic feature of longitudinal data is the statistical dependence between the repeated measurements for a subject, which has to be addressed to quantify the association. Linear mixed effects models (Laird et al., 1982) provide one way to model the repeated measurements and are widely-used in behavioral sciences (Gibbons et al., 2010). The model parameters can be interpreted as the desired association. Hence, the design problem is to choose the time points for taking the measurements in order to maximize statistical efficiency of the parameter estimates. The problem is of practical importance particularly for mobile phone-based studies of human behavior (Shiffman et al., 2008; Kirchner et al., 2013; Kaplan and Stone, 2013; Smets et al., 2018; Wang et al., 2014) which collect intensive data through self-reports and mobile sensors. The use of machine learning combined with optimal design methods, as demonstrated in our approach, has the potential to improve the inferences drawn from such studies.

Related Work Designs for longitudinal studies have primarily focused on linear mixed effects models (Sinha and Xu, 2011, 2016; Liu et al., 2012). Finding optimal measurement times for analysis via such models has also been studied (Winkens et al., 2005; Berger and Tan, 2004; Ouwens et al., 2002). Unfortunately, the design criterion depends on the future covariate values. These works only consider covariates that are known functions of time or are known beforehand, which is a commonly used setup (e.g. see Fedorov and Hackl, 1997; Atkinson et al., 2007, Chapter 24). Hence, the covariates and the designs can be computed beforehand. Literature on constrained optimal designs considers the case when some covariates are not under experimenter's control and are not known before the experiment. It is assumed that marginal distribution of the uncontrolled covariates or their distribution conditioned on the controlled covariates is known beforehand, referred to as marginally or conditionally restricted designs (Cook and Thibodeau, 1980; Lopez-Fidalgo and Garcet-Rodríguez, 2004), respectively. Optimal designs are searched in a restricted set that adheres to these distributions for the uncontrolled covariates. In contrast, we consider the case where covariate distribution is not known.

Current Work We focus on constructing optimal designs for linear mixed effects model. In contrast to prior work, the future covariates are modelled by an unknown function of time and past covariates. We use a flexible autoregressive model to represent the function and learn it sequentially as more data is observed. Estimated future covariate values are then used to compute designs optimizing the criterion value.

2. Background

We will consider linear mixed effects model with continuous-valued responses. Next, we describe the model assumptions, followed by the sequential design procedure for the model.

Linear Mixed Effects Model Suppose, individual $i \in \{1, 2, \ldots, N\}$ has m measurements in time. Then, the $m \times 1$ vector of responses Y_i is assumed to follow,

$$Y_i = \mathbf{X}_i \beta + \mathbf{Z}_i b_i + \mathbf{e}_i. \tag{1}$$

Here, $\mathbf{X}_i$ is an $m \times p$ matrix of covariates at m measurement times, including both time-fixed and time-varying variables, and covariates $\mathbf{Z}_i \subset \mathbf{X}_i$. Elements of β give the associations of interest for the study, while b_i are considered nuisance parameters. The term $\mathbf{e}_i \sim N(\mathbf{0}, \sigma^2 \mathbf{V}_i)$ is considered random error which can be modeled in different ways, including as $\mathbf{V}_i = \mathbf{I}$ or as an autoregressive process. Further, it is assumed that $b_i \sim N(0, \mathbf{\Psi}), b_i \perp\!\!\!\perp \mathbf{X}_i$, and $b_i \perp\!\!\!\perp \mathbf{e}_i$. Under these assumptions, we obtain $Y_i \sim N(\mathbf{X}_i\beta, \mathbf{W}_i)$ where $\mathbf{W}_i = \mathbf{Z}_i \mathbf{\Psi} \mathbf{Z}_i^\top + \sigma^2 \mathbf{V}_i$ (Schafer, 2006, Chapter 1 p. 10). Thus, introducing the random effects b_i helps to model the dependence between measurements in time. The maximum likelihood estimate $\hat{\beta}$ for β is obtained by solving the following estimating equation,

$$\frac{1}{N} \sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{W}_i^{-1} (Y_i - \mathbf{X}_i \beta) = 0, \tag{2}$$

and the variance-covariance (in brief, the variance) matrix of the resulting estimator (Berger and Tan, 2004) is given by,

$$\mathrm{var}(\hat{\beta}) := \left[\sum_{i=1}^{N} (\mathbf{X}_i^\top \mathbf{W}_i^{-1} \mathbf{X}_i) \right]^{-1} \tag{3}$$

Optimal Design of Experiments Design of experiments framework requires defining a design criterion to evaluate the utility of collecting data according to a design. Different design criteria have been proposed in the literature, based on minimizing different functions of the variance matrix, thereby increasing statistical efficiency of the estimate. A popular method, termed as D-optimal design, minimizes the determinant of the variance which, geometrically, represents the volume of the confidence region of the estimate. Other criteria such as A-optimal (trace of the variance matrix) and C-optimal (variance of $c^\top \hat{\beta}$ for a vector c) are also used. The optimal designs are defined as solutions to the following optimization problem,

$$\arg\min_{\tau \in \mathcal{T}} \phi\big(\mathrm{var}(\hat{\beta}_\tau)\big). \tag{4}$$

where $\mathcal{T}$ denotes the design space, $\hat{\beta}_\tau$ is the estimate from data collected by design τ, and $\phi(A)$ is the design criterion chosen by the experimenter, such as $\phi(A) = \det(A)$, the determinant of the variance matrix A. In our case, a design $\tau \in \mathcal{T}$ denotes m_i measurement times for each individual $i \in \{1, 2, \ldots, N\}$. Thus, $\tau := (\tau_1, \tau_2, \ldots, \tau_N)$ where each $\tau_i \in [0, T]^{m_i}$, for some pre-specified study period T.

For the linear mixed effects model, the criterion (3) depends on the values of unknown parameters $\mathbf{\Psi}, \sigma^2, \mathbf{V}_i$, and the values of covariates $\mathbf{X}_i$ measured at arbitrary times τ_i. The unknown parameters can be estimated sequentially based on previous measurements which can be used to find a (locally) optimal design. However, the covariate values at future time points are still required. Earlier work (Berger and Tan, 2004; Sinha and Xu, 2011) assumes that the covariates are only polynomial functions of time. Thus, the design criterion can be evaluated for arbitrary measurement times. The main idea in our work is to use a forecasting model to predict the covariate values from past measurements and using these to approximately minimize the design criterion.

Algorithm 1: Finding optimal measurement times

```
Input: Total study period $T$ days, Forecasting model, Association model
Output: Measurement times, Parameter estimates
Initialize estimates for model parameters (1), possibly, using some past data;
for $t \leftarrow 1$ to $T$ days do
    1. Construct matrix $\mathbf{X}_i$ for time points in day $t$ using the forecasting model;
    2. Calculate $\mathbf{W}_i$ using $\mathbf{X}_i$ and current estimates of model parameters;
    3. Select the design minimizing (4) from the time points in day $t$;
    4. Sample covariates and responses for the selected time points for each
       individual;
    5. Update forecasting model and association model;
end
Return parameter estimates for the association model;
```

3. Approach

The process of computing designs is guided by two models – a *forecasting model* to extrapolate covariate values and an *association model* to evaluate variance of the association estimates for a given design.

Forecasting Model Suppose, the observed data till time P across all individuals is denoted by $\mathcal{D}_P := \{(\mathbf{X}_i(t), Y_i(t)) \mid t \in M_i, t \leq P, i \in \{1, 2, \ldots, N\}\}$, where M_i is the set of measured times for individual i. The task is to construct a prediction function $f(t, i; \mathcal{D}_P) = \mathbb{E}(\mathbf{X}_i(t) \mid \mathcal{D}_P)$ for any future time $t > P$ and any individual $i \in \{1, 2, \ldots, N\}$. Any flexible autoregressive model can be used. Unlike the association model, we are not required to make inferences on the parameters of the forecasting model. We use the Gaussian Process Autoregressive model (GPAR; Requeima et al., 2019), motivated by the ability to express assumptions about covariate evolution through its kernel function. The input features for the model are time t and one-hot encoding of the individual index i. By adding the individual index as input, we can train a single model for all individuals and still distinguish among observations from different individuals in the input representation.

Association Model We consider the linear mixed effects model (1) for estimating association, given by β. Restricted maximum likelihood-based estimation is used for parameters in the random effects and noise components i.e. $\Psi, \sigma^2, \mathbf{V}_i$ (Schafer, 2006) and $\hat{\beta}$ is obtained by the ordinary least squares estimate (2).

The steps involved in the design process are sketched in Algorithm 1. The study period is divided into time blocks, say by days. For each time block, a design is computed by minimizing the design criterion, followed by sampling data according to the design. Finally, both the models are updated for use in the next time block.

Next, we describe a simulation setting to test the feasibility of the approach. We specify the data generating process, the heuristic used for optimization, and the evaluation metrics.

4. Experiments

To understand the behavior of the proposed approach, we start with experiments on synthetic datasets where the underlying data generating process is known.

Synthetic Data For each individual $i \in \{1, 2, \ldots, 20\}$ and for 100 equally-spaced time points $t \in [0, 1]$, covariates and responses are generated from a linear mixed effects model as follows,

$$\begin{aligned} \mathbf{X}_i(t) &= [1 \quad x_1(t) \quad x_2(t)], \quad \mathbf{Z}_i(t) = [1], \\ b_i &\sim \mathcal{N}(0, 0.1), \\ Y_i(t) &\sim \mathcal{N}(\mathbf{X}_i(t)\beta + \mathbf{Z}_i(t)b_i, \sigma^2), \\ \beta &= [1 \quad -0.5 \quad 1]^\top, \sigma^2 = 0.1 \end{aligned}$$

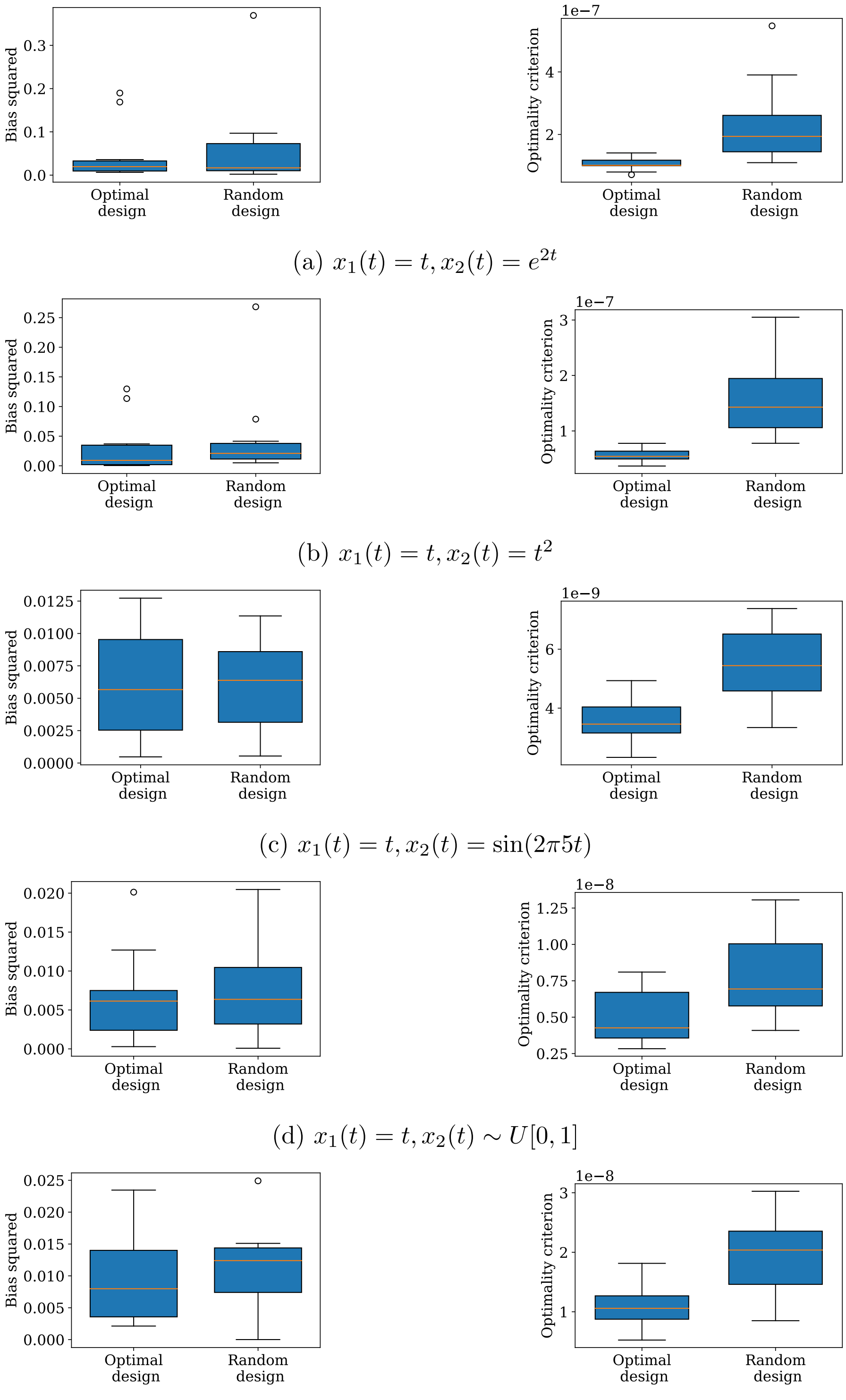
Covariates $(x_1(t), x_2(t))$ are constructed using different functions such as sinusoidal and exponential. Design criterion is taken to be D-optimality, i.e. the objective function in (4) is $\det(\text{var}(\hat{\beta}_\mathcal{T}))$. To solve the optimization problem, time range $[0, 1]$ is discretized into 100 time points. Then, we use greedy search to select m time points i.e. each of the m iterations selects the time point with minimum objective function value when combined with already selected points. To simplify the search space and to reduce computations, we assume that all individuals are sampled at the same design points. More principled search procedures such as Federov's exchange algorithm (Cook and Nachtrheim, 1980) can be investigated. First 50 time points for all $N = 20$ individuals are considered to be fully observed. The next 50 time points are the test set, where we select $m = 10$ time points according to the design criterion using forecasting and association models learned from the training data.

Evaluation Measurement times obtained using the optimal design framework are compared with times chosen with uniform random sampling, in which $m = 10$ time points are sampled uniformly at random from the 50 test points. Since we know the true parameters β for the synthetic data, we can compute the bias in estimates $\hat{\beta}$, quantified as bias squared $\|\beta - \hat{\beta}\|_2^2$. Variance of $\hat{\beta}$ is computed empirically over multiple runs with sampled datasets.

Results The estimates from the two design criteria with different covariate functions are compared in Figure 1. Comparing the bias squared, we observe that optimal designs have similar or slightly lower bias than uniform random designs (Figure 1, first column). We also plot the design criterion, i.e. $\det\big(\text{var}(\hat{\beta})\big)$, for both the designs, observing that D-optimal designs have lower values as expected (Figure 1, second column). Thus, the proposed approach increases the statistical efficiency of estimation while keeping the bias low.

5. Conclusion and Future Work

We describe the problem of finding optimal designs for longitudinal data analysis where design space is not fully observable or controllable. Advances on the problem can lead to improved designs for mobile phone-aided studies of human behavior. Through simulations, we observe promising preliminary results for the proposed approach that employs machine learning for predicting the unobserved covariates. Future work includes study of designs under larger model classes e.g. under non-linear mixed effects models (Foster et al., 2018) and a principled way to incorporate prediction uncertainty in the design process.



(a) $x_1(t) = t, x_2(t) = e^{2t}$

(b) $x_1(t) = t, x_2(t) = t^2$

(c) $x_1(t) = t, x_2(t) = \sin(2\pi 5t)$

(d) $x_1(t) = t, x_2(t) \sim U[0, 1]$

(e) $x_1(t) = t, x_2(t) \sim \mathcal{N}(0, 1)$

Figure 1: Bias squared (first column) and design criterion value (second column) for estimates by D-optimal designs and uniform random sampling-based designs. Box plots show the mean along with first and third quantiles from 10 simulated datasets. Observe that optimal designs result in low bias while significantly decreasing the variance of the estimate.

References

Anthony Atkinson, Alexander Donev, and Randall Tobias. *Optimum experimental designs, with SAS*, volume 34. Oxford University Press, 2007.

Martijn PF Berger and Frans ES Tan. Robust designs for linear mixed effects models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(4):569–581, 2004.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

R Dennis Cook and Christopher J Nachtrheim. A comparison of algorithms for constructing exact d-optimal designs. *Technometrics*, 22(3):315–324, 1980.

R Dennis Cook and LA Thibodeau. Marginally restricted d-optimal designs. *Journal of the American Statistical Association*, 75(370):366–371, 1980.

Valerii Fedorov. Optimal experimental design. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(5):581–589, 2010.

Valerii V. Fedorov and Peter Hackl. *Special Cases and Applications*, pages 69–105. Springer New York, New York, NY, 1997. ISBN 978-1-4612-0703-0. doi: 10.1007/978-1-4612-0703-0_6. URL `https://doi.org/10.1007/978-1-4612-0703-0_6`.

Adam Foster, Martin Jankowiak, Eli Bingham, Yee Whye Teh, Tom Rainforth, and Noah Goodman. Variational optimal experiment design: Efficient automation of adaptive experiments. *NeurIPS Workshop on Bayesian Deep Learning*, 2018. URL `http://bayesiandeeplearning.org/2018/papers/90.pdf`.

Robert D Gibbons, Donald Hedeker, and Stephen DuToit. Advances in analysis of longitudinal data. *Annual review of clinical psychology*, 6:79–107, 2010.

Kirthevasan Kandasamy, Willie Neiswanger, Reed Zhang, Akshay Krishnamurthy, Jeff Schneider, and Barnabas Poczos. Myopic posterior sampling for adaptive goal oriented design of experiments. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3222–3232, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL `http://proceedings.mlr.press/v97/kandasamy19a.html`.

Robert M. Kaplan and Arthur A. Stone. Bringing the laboratory and clinic to the community: Mobile technologies for health promotion and disease prevention. *Annual Review of Psychology*, 64(1):471–498, 2013. doi: 10.1146/annurev-psych-113011-143736. URL `https://doi.org/10.1146/annurev-psych-113011-143736`. PMID: 22994919.

Darla E Kendzor, Kerem Shuval, Kelley Pettee Gabriel, Michael S Businelle, Ping Ma, Robin R High, Erica L Cuate, Insiya B Poonawalla, Debra M Rios, Wendy Demark-Wahnefried, et al. Impact of a mobile phone intervention to reduce sedentary behavior in a community sample of adults: a quasi-experimental evaluation. *Journal of medical Internet research*, 18(1):e19, 2016.

Thomas R Kirchner, Jennifer Cantrell, Andrew Anesetti-Rothermel, Ollie Ganz, Donna M Vallone, and David B Abrams. Geospatial exposure to point-of-sale tobacco: real-time craving and smoking-cessation outcomes. *American journal of preventive medicine*, 45 (4):379–385, 2013.

Nan M Laird, James H Ware, et al. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982.

Qing Liu, Angela M Dean, and Greg M Allenby. Bayesian designs for hierarchical linear models. *Statistica Sinica*, pages 393–417, 2012.

Jesus Lopez-Fidalgo and Sandra A Garcet-Rodríguez. Optimal experimental designs when some independent variables are not subject to control. *Journal of the American Statistical Association*, 99(468):1190–1199, 2004.

Mario JNM Ouwens, Prans ES Tan, and Martijn PF Berger. Maximin d-optimal designs for longitudinal mixed effects models. *Biometrics*, 58(4):735–741, 2002.

James Requeima, William Tebbutt, Wessel Bruinsma, and Richard E Turner. The gaussian process autoregressive regression model (gpar). In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1860–1869, 2019.

Philip A Romero, Andreas Krause, and Frances H Arnold. Navigating the protein fitness landscape with gaussian processes. *Proceedings of the National Academy of Sciences*, 110 (3):E193–E201, 2013.

Theodore A Walls Joseph L Schafer. *Models for intensive longitudinal data*. Oxford University Press, 2006.

Saul Shiffman, Arthur A Stone, and Michael R Hufford. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.*, 4:1–32, 2008.

Sanjoy K Sinha and Xiaojian Xu. Sequential d-optimal designs for generalized linear mixed models. *Journal of Statistical Planning and Inference*, 141(4):1394–1402, 2011.

Sanjoy K Sinha and Xiaojian Xu. Sequential designs for repeated-measures experiments. *Journal of Statistical Theory and Practice*, 10(3):497–514, 2016.

Elena Smets, Emmanuel Rios Velazquez, Giuseppina Schiavone, Imen Chakroun, Ellie D'Hondt, Walter De Raedt, Jan Cornelis, Olivier Janssens, Sofie Van Hoecke, Stephan Claes, et al. Large-scale wearable data reveal digital phenotypes for daily-life stress detection. *npj Digital Medicine*, 1(1):67, 2018.

Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*, pages 3–14. ACM, 2014.

Bjorn Winkens, Hubert JA Schouten, Gerard JP van Breukelen, and Martijn PF Berger. Optimal time-points in clinical trials with linearly divergent treatment effects. *Statistics in medicine*, 24(24):3743–3756, 2005.