Sankhyā : The Indian Journal of Statistics
2021, Volume 83-A, Part 2, pp. 745-778

Optimal Designs for Model Averaging in non-nested Models

Kira Alhorn and Kirsten Schorning
Technische Universität Dortmund, Dortmund, Germany
Holger Dette
Ruhr-Universität Bochum, Bochum, Germany

Abstract

In this paper we construct optimal designs for frequentist model averaging estimation. We derive the asymptotic distribution of the model averaging estimate with fixed weights in the case where the competing models are non-nested. A Bayesian optimal design minimizes an expectation of the asymptotic mean squared error of the model averaging estimate calculated with respect to a suitable prior distribution. We derive a necessary condition for the optimality of a given design with respect to this new criterion. We demonstrate that Bayesian optimal designs can improve the accuracy of model averaging substantially. Moreover, the derived designs also improve the accuracy of estimation in a model selected by model selection and model averaging estimates with random weights.

AMS (2000) *subject classification.* Primary 62K05; Secondary 62F12.
Keywords and phrases. Model selection, Model averaging, Uniform weighting, Model uncertainty, Optimal design, Bayesian optimal design.

1 Introduction

There exists an enormous amount of literature on selecting an adequate model from a set of candidate models for statistical analysis. Numerous model selection criteria have been developed for this purpose. These procedures are widely used in practice and have the advantage of delivering a single model from a class of competing models, which makes them very attractive for practitioners. Exemplarily, we mention Akaike's information criterion (AIC), the Bayesian information criterion (BIC) and its extensions,

The author gratefully acknowledges financial support by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823, Teilprojekt C2, T1) of the German Research Foundation (DFG).

Mallow's C_p, the generalized cross-validation and the minimum description length (see the monographs of Burnham and Anderson (2002) & Konishi and Kitagawa (2008) and Claeskens and Hjort (2008) for more details). Different criteria have different properties, such as consistency, efficiency, minimax-rate optimality, and parsimony (used in the sense of Claeskens and Hjort, (2008, Chapter 4).

Yang (2005) proves that the different properties cannot be combined and that there is no universally optimal model selection criterion in sense of consistency and minimax-optimality. Consequently, different criteria might be preferable in different situations depending on the particular application.

On the other hand, there exists a well known post-selection problem in this approach because model selection introduces an additional variance that is often ignored in statistical inference after model selection (see Pötscher (1991) for one of the first contributions discussing this issue). This post-selection problem is inter alia attributable to the fact, that estimates after model selection behave like mixtures of potential estimates. For example, ignoring the model selection step (and thus the additional variability) may lead to confidence intervals with coverage probability smaller than the nominal value, see for example Chapter 7 in Claeskens and Hjort (2008) for a mathematical treatment of this phenomenon.

An alternative to model selection is model averaging, where estimates of a target parameter are smoothed across several models, rather than restricting inference on a single selected model.

This approach has been widely discussed in the Bayesian literature, where it is known as "Bayesian model averaging" (see the tutorial of Hoeting et al. (1999) among many others). For Bayesian model averaging prior probabilities have to be specified. This might not always be possible and therefore Buckland et al. (1997) proposed a "frequentist model averaging", where smoothing across several models is commonly based on information criteria. Kapetanios et al. (2008) demonstrated that the frequentist approach is a worthwhile alternative to Bayesian model averaging. Stock and Watson (2003) observed that averaging predictions usually performs better than forecasting in a single model. Hong and Preston (2012) substantiate these observations with theoretical findings for Bayesian model averaging if the competing models are "sufficiently close". Further results pointing in this direction can be found in Raftery and Zheng (2003) & Schorning et al. (2016) and Buatois et al. (2018).

Independently of this discussion there exists a large amount of research how to optimally design experiments under model uncertainty (see Box and Hill (1967) & Atkinson and Fedorov (1975) for early contributions). This

work is motivated by the fact that an optimal design can improve the efficiency of the statistical analysis substantially if the postulated model assumptions are correct, but may be inefficient if the model is misspecified. Many authors suggested to choose the design for model discrimination such that the power of a test between competing regression models is maximized (see Ucinski and Bogacka (2005), López-Fidalgo et al. (2007), & Tommasi and López-Fidalgo (2010) or Dette et al. (2015) for some more recent references). Other authors proposed to minimize an average of optimality criteria from different models to obtain an efficient design for all models under consideration (see Dette (1990), Zen and Tsai (2002), & Tommasi (2009) among many others).

Although model selection or averaging are commonly used tools for statistical inference under model uncertainty most of the literature on designing experiments under model uncertainty does not address the specific aspects of these methods directly. Optimal designs are usually constructed to maximize the power of a test for discriminating between competing models or to minimize a functional of the asymptotic variance of estimates in the different models. To the best of our knowledge (Alhorn et al., 2019) is the first contribution, which addresses the specific challenges of designing experiments for model selection or model averaging. These authors constructed optimal designs minimizing the asymptotic mean squared error of the model averaging estimate and showed that optimal designs can yield a reduction of the mean squared error up to 45%. Moreover, they also showed that these designs improve the performance of estimates in models chosen by model selection criteria. However, their theory relies heavily on the assumption of nested models embedded in a framework of local alternatives as developed by Hjort and Claeskens (2003).

The goal of the present contribution is the construction of optimal designs for model averaging in cases where the competing models are not nested (note that in this case local alternatives cannot be formulated).

In order to derive an optimality criterion, which can be used for the determination of optimal designs in this context, we further develop the approach of Hjort and Claeskens (2003) and derive an asymptotic theory for model averaging estimates for classes of competing models which are non-nested. Optimal designs are then constructed minimizing the asymptotic mean squared error of the model averaging estimate and it is demonstrated that these designs yield substantially more precise model averaging estimates. Moreover, these designs also improve the performance of estimates after model selection. Our work also contributes to the discussion of the superiority of model averaging over model selection. Most of the results presented in literature

indicate that model averaging has some advantages over model selection in general. We demonstrate that conclusions of this type depend sensitively on the class of models under consideration. In particular, we observe some advantages of estimation after model selection if the competing models are of rather different shape for small sample sizes. Nevertheless, the optimal designs developed in this paper improve both estimation methods, where the improvement can be substantial in many cases.

The remaining part of this paper is organized as follows. The pros and cons of model averaging and model selection are briefly discussed in Section 2 where we introduce the basic methodology and investigate the impact of similarity of the candidate models on the performance of the different estimates. In Section 3 we develop asymptotic theory for model averaging estimation in the case where the models are non-nested. Based on these results we derive a criterion for the determination of optimal designs and establish a necessary condition for optimality. In Section 4 we study the performance of these designs by means of a simulation study. In Section 5 we discuss some robustness properties of the optimal designs if either the true data generating model is not contained in the set of competing models or if they are used to estimate other parameters. Finally, technical assumptions and proofs are given Appendix.

2 Model Averaging Versus Model Selection

In this section we introduce the basic terminology and also illustrate in a regression framework that the superiority of model averaging about estimation in a model chosen by model selection depends sensitively on the class of competing models.

2.1. Basic Terminology We consider data obtained at k different experimental conditions, say $x_1, \ldots, x_k$ chosen in a design space $\mathcal{X}$. At each experimental condition x_i one observes n_i responses, say $y_{i1}, \ldots, y_{in_i}$ $(i = 1, \ldots, k)$, and the total sample size is $n = \sum_{i=1}^{k} n_i$. We also assume that the responses $y_{i1}, \ldots, y_{in_i}$ are realizations of random variables of the form

$$Y_{ij} = \eta_s(x_i, \vartheta_s) + \varepsilon_{ij}, i = 1, \ldots, k, j = 1, \ldots, n_i, s = 1, \ldots, r, \tag{2.1}$$

where the regression function η_s is a differentiable function with respect to the parameter ϑ_s and the random errors ε_{ij} are independent normally distributed with mean 0 and common variance σ^2. Furthermore, the index s in η_s corresponds to different models (with parameters ϑ_s) and we assume that there are r competing regression functions $\eta_1, \ldots, \eta_r$ under consideration.

Having r different candidate models (differing by the regression functions η_s) a classical approach for estimating a parameter of interest, say μ, is to

calculate an information criterion for each model under consideration and estimate this parameter in the model optimizing this criterion. For this purpose, we denote the density of the normal distribution corresponding to a regression model (2.1) by $f_s(\cdot \mid x_i, \theta_s)$ with parameter $\theta_s = (\sigma^2, \vartheta_s)^\top$ and identify the different models by their densities $f_1, \ldots, f_r$ (note that in the situation considered in this sections these only differ in the mean). Using the observations $y_n = (y_{11}, \ldots, y_{1n_1}, y_{21}, \ldots, y_{kn_k})^\top$ we calculate in each model the maximum likelihood estimate

$$\hat{\theta}_{n,s} = \arg\max_{\theta_s \in \Theta_s} \ell_{n,s}(\theta_s \mid y_n) \tag{2.2}$$

of the parameter θ_s, where

$$\ell_{n,s}(\theta_s \mid y_n) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \log f_s(y_{ij} \mid x_i, \theta_s) \tag{2.3}$$

is the log-likelihood in candidate model f_s $(s = 1, \ldots r)$. Each estimate $\hat{\theta}_{n,s}$ of the parameter θ_s yields an estimate $\hat{\mu}_s = \mu_s(\hat{\theta}_{n,s})$ for the quantity of interest, where μ_s is the target parameter in model s.

For example, regression models of the type (2.1) are frequently used in dose finding studies (see MacDougall (2006) or Bretz et al., 2008). In this case a typical target function μ_s of interest is the "quantile" defined by

$$\mu_s(\theta_s) = \inf \left\{ x \in \mathcal{X} \;\middle|\; \frac{\eta_s(x, \vartheta_s) - \eta_s(a, \vartheta_s)}{\eta_s(b, \vartheta_s) - \eta_s(a, \vartheta_s)} \geq \alpha \right\}. \tag{2.4}$$

The value defined in Eq. 2.4 is well-known as ED_α, that is, the effective dose at which $100 \times \alpha\%$ of the maximum effect in the design space $\mathcal{X} = [a, b]$ is achieved.

We now briefly discuss the principle of model selection and averaging to estimate the target parameter μ. For model selection we choose the model f_{s^*} from $f_1, \ldots, f_s$, which maximizes Akaike's information criterion (AIC)

$$\mathrm{AIC}(f_s \mid y_n) = \frac{2}{n} \ell_{n,s}(\hat{\theta}_{n,s} \mid y_n) - 2p_s, \tag{2.5}$$

where p_s is the number of parameters in model f_s (see Claeskens and Hjort, 2008, Chapter 2). The target parameter is finally estimated by $\hat{\mu} = \mu_{s^*}(\hat{\theta}_{n,s^*})$. Obviously, other model selection schemes, such as the Bayesian or focussed information criterion can be used here as well, but we restrict ourselves to the AIC for the sake of a transparent presentation.

Roughly speaking, model averaging is a weighted average of the individual estimates in the competing models. It might be viewed from a Bayesian (see for example Wassermann, 2000) or a frequentist point of view (see for example Claeskens and Hjort, 2008) resulting in different choices of model averaging weights. We will focus here on non-Bayesian methods. More explicitly, assigning nonnegative weights $w_1, \ldots, w_r$ to the candidate models $f_1, \ldots, f_r$, with $\sum_{i=1}^{r} w_i = 1$, the model averaging estimate for μ is given by

$$\hat{\mu}_{\text{mav}} = \sum_{s=1}^{r} w_s \mu_s(\hat{\theta}_{n,s}). \tag{2.6}$$

Frequently used weights are uniform weights (see, for example Stock and Watson (2004) & Kapetanios et al. (2008)). More elaborate model averaging weights can be chosen depending on the data. For example, Buckland et al. (1997) define smooth AIC-weights as

$$w_s^{\text{smAIC}}(y_n) = \frac{\exp\{\frac{1}{2}\text{AIC}(f_s \mid y_n)\}}{\sum_{s=1}^{r} \exp\{\frac{1}{2}\text{AIC}(f_s \mid y_n)\}}. \tag{2.7}$$

Alternative data dependent weights can be constructed using other information criteria or model selection criteria. There also exists a vast amount of literature on determining optimal data dependent weights such that the resulting mean squared error of the model averaging estimate is minimal (see Hjort and Claeskens (2003), Hansen (2007), & Zhang et al. (2016) and Liang et al. (2011) among many others). For the sake of brevity, we will concentrate on smooth AIC-weights, which are frequently used in the context of dose-finding studies (see Sébastien et al. (2016) & Verrier et al. (2014), among others). Nevertheless, similar observations as presented in this paper could be made for other data dependent weights which are constructed using information criteria like the Bayesian information criterion.

2.2. The Class of Competing Models Matters In this section we illustrate the influence of the candidate set on the properties of model averaging estimation and estimation after model selection by means of a brief simulation study. For this purpose we consider four regression models of the form (2.1), which are commonly used in dose-response modeling and specified in Table 1 with corresponding parameters.

Here we adapt the setting of Pinheiro et al. (2006) who model the dose-response relationship of an anti-anxiety drug, where the dose of the drug may vary in the interval $\mathcal{X} = [0, 150]$. In particular, we have $k = 6$ different dose levels $x_i \in \{0, 10, 25, 50, 100, 150\}$ and patients are allocated to each dose level most equally, where the total sample size is $n \in \{50, 100, 250\}$.

Table 1: Models and parameters used for the simulation study

Model	Mean function η_s	Parameter specifications
Log-Linear (f_1)	$\eta_1(x_i, \vartheta_1) = \vartheta_{11} + \vartheta_{12} \log(x_i + \vartheta_{13})$	$\vartheta_1 = (0, 0.0797, 1)^\top$
Emax (f_2)	$\eta_2(x_i, \vartheta_2) = \vartheta_{21} + \frac{\vartheta_{22} x}{\vartheta_{23} + x}$	$\vartheta_2 = (0, 0.467, 25)^\top$
Exponential (f_3)	$\eta_3(x_i, \vartheta_3) = \vartheta_{31} + \vartheta_{32} \exp(x_i / \vartheta_{33})$	$\vartheta_3 = (-0.08265, 0.08265, 85)^\top$
Quadratic (f_4)	$\eta_4(x_i, \vartheta_4) = \vartheta_{41} + \vartheta_{42} x + \vartheta_{43} x^2$	$\vartheta_4 = (0, 0.00533, -0.00002)^\top$

Additionally, we present results for a larger sample size ($n = 1000$) in order to investigate the asymptotic properties of the different estimation methods.

We consider the problem of estimating the $\mathrm{ED}_{0.4}$, as defined in Eq. 2.4.

To investigate the particular differences between both estimation methods we choose two different sets of competing models from Table 1. The first set

$$\mathcal{S}_1 = \{f_1, f_2, f_4\} \tag{2.8}$$

contains the log-linear, the Emax and the quadratic model, while the second set

$$\mathcal{S}_2 = \{f_1, f_2, f_3\} \tag{2.9}$$

contains the log-linear, the Emax and the exponential model. The set $\mathcal{S}_1$ serves as a prototype set of "similar" models while the set $\mathcal{S}_2$ contains models of more "different" shape. This is illustrated in Fig. 1. In the left panel we show the quadratic model f_4 (for the parameters specified in Table 1) and the best approximations of this function by a log-linear model (f_1) and an Emax model (f_2) with respect to the Kullback-Leibler divergence

$$\frac{1}{6}\sum_{i=1}^{6}\int f_4(y \mid x_i, \theta_4)\log\left(\frac{f_4(y \mid x_i, \theta_4)}{f_s(y \mid x_i, \theta_s)}\right)dy,\ \ s = 1, 2. \tag{2.10}$$

In this case, all models have a very similar shape and we obtain for the $\mathrm{ED}_{0.4}$ the values 32.581, 32.261 and 33.810 for the log-linear (f_1), Emax (f_2) and quadratic model (f_4). Similarly the right panel shows the exponential model (f_3, solid line) and its corresponding best approximations by the log-linear model (f_1) and the Emax model (f_2). Here we observe larger differences between the models in the candidate set and we obtain for the $\mathrm{ED}_{0.4}$ the values 58.116, 42.857 and 91.547 for the models f_1, f_2 and f_3, respectively.

All results presented in this paper are based on 1000 simulation runs generating in each run n observations of the form

$$y_{ij}^{(l)} = \eta_s(x_i, \vartheta_s) + \varepsilon_{ij}^{(l)}, i = 1, \ldots, k, j = 1, \ldots, n_i, \tag{2.11}$$

where the errors $\varepsilon_{ij}^{(l)}$ are independent centered normal distributed random variables with $\sigma^2 = 0.1$ and η_s is one of the models $\eta_1, \ldots, \eta_4$ (with parameters specified in Table 1). The parameter $\mu = \mathrm{ED}_{0.4}$ is estimated by model averaging with uniform weights, smooth AIC-weights given in Eq. 2.7 and estimation after model selection by the AIC.

In Tables 2 and 3 we show the simulated mean squared errors of the model averaging estimates with uniform weights (left column), smooth AIC-weights

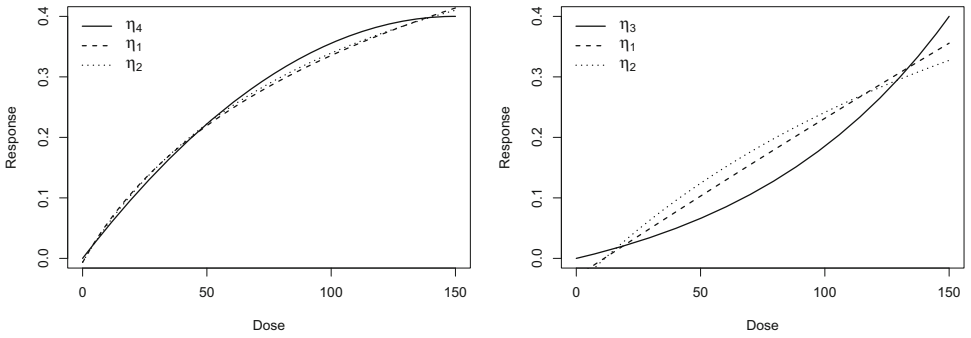

Figure 1: Left panel: quadratic model (solid line) and its best approximations by the log-linear (dashed line) and the Emax model (dotted line) with respect to the Kullback-Leibler divergence (2.10). Right panel: exponential model (solid line) and its best approximations by the log-linear (dashed line) and the Emax model (dotted line)

(middle column) and estimation after model selection (right column). Here, different rows correspond to different models. The numbers printed in bold face indicate the estimation method with the smallest mean squared error.

Table 2: Simulated mean squared error of different estimates of the $ED_{0.4}$

Model	Sample size	Uniform weights	Smooth AIC-weights	Model selection
	$n = 50$	**437.045**	498.323	758.978
f_1	$n = 100$	223.291	**218.99**	285.062
	$n = 250$	111.973	82.713	**78.371**
	$n = 1000$	58.183	9.488	**9.022**
	$n = 50$	**286.638**	329.904	515.32
f_2	$n = 100$	**189.785**	203.796	251.836
	$n = 250$	**62.792**	64.854	66.54
	$n = 1000$	23.150	**13.922**	14.651
	$n = 50$	**276.037**	361.101	669.873
f_4	$n = 100$	**190.662**	244.558	391.443
	$n = 250$	**92.653**	109.852	139.859
	$n = 1000$	26.370	**26.166**	26.244

The set of candidate models is $\mathcal{S}_1 = \{f_1, f_2, f_4\}$. Left column: model averaging with uniform weights; middle column: model averaging with smooth AIC-weights; right column: estimation after model selection

Table 3: Simulated mean squared error of different estimates of the $ED_{0.4}$

Model	Sample size	Estimation method: Uniform weights	Smooth AIC-weights	Model selection
f_1	$n = 50$	834.295	**553.427**	776.311
	$n = 100$	712.404	**340.254**	353.707
	$n = 250$	524.518	48.587	**38.591**
	$n = 1000$	479.261	7.977	**7.957**
f_2	$n = 50$	640.706	**505.054**	669.285
	$n = 100$	517.963	**267.967**	286.272
	$n = 250$	394.536	65.805	**53.424**
	$n = 1000$	348.071	10.046	**9.899**
f_3	$n = 50$	**1076.154**	1141.476	1427.441
	$n = 100$	871.362	**766.140**	802.763
	$n = 250$	802.196	480.641	**399.839**
	$n = 1000$	758.983	99.948	**85.171**

The set of candidate models is $\mathcal{S}_2 = \{f_1, f_2, f_3\}$. Left column: model averaging with uniform weights; middle column: model averaging with smooth AIC-weights; right column: estimation after model selection

2.2.1. Models of similar shape. We will first discuss the results for the set of similar models in Eq. 2.8 (see Table 2). If the sample size is small, model averaging with uniform weights performs very well. Model averaging with smooth AIC-weights yields an about 10% -25% larger mean squared error (except for two cases, where it performs better than model averaging with uniform weights). On the other hand the mean squared error of estimation after model selection is substantially larger than that of model averaging, if the sample size is small. This is a consequence of the additional variability associated with data-dependent weights. For example, if the sample size is $n = 50$ and the data generating model is given by f_1, the mean squared errors of the model averaging estimates with uniform and smooth AIC-weights and the estimate after model selection are given by 437.0, 498.3 and 759.0, respectively. The corresponding variances are given by 235.2, 337.6 and 599.7, respectively. For the squared bias the order is exactly the opposite, that is 201.9, 160.7, 159.3, but the differences are not so large. This means that the bias can be reduced by using random weights, because these put more weight on the "correct" model.

If the sample size is $n = 1000$, the mean squared error of the model averaging estimates with uniform weights is larger than the mean squared errors

obtained by smooth AIC-weights and the estimate after model selection. Exemplarily, if the data generating model is given by f_1, the model averaging estimator with uniform weights yields an about five times larger mean squared error than the model averaging estimator with smooth AIC-weights and the estimate after model selection. Thus, both AIC-based methods outperform the model averaging estimate with uniform weights if the sample size is large. This behaviour can be explained by the fact that the AIC is weakly consistent, i.e. in the setting under consideration, it selects the best (true) model with probability converging to one with increasing sample size. Consequently, for large sample size the model averaging estimator and the estimator after model selection do not differ much and either often select or put a high weight to the true data generating model.

Summarizing, for small sample sizes model averaging performs better than estimation after model selection. These observations coincide with the findings of Schorning et al. (2016) and Buatois et al. (2018) who compared model averaging and model selection in the context of dose finding studies (see also Chen et al. (2018) for similar results for the AIC in the context of ordered probit and nested logit models). In particular, model averaging with (fixed) uniform weights yields very reasonable results in our case. Note that the phenomenon that model averaging with uniform weights can improve the estimation accuracy in comparison to the estimation after model selection and even outperforms other averaging methods can also be observed in other situations (see, for example, Bates and Granger (1969) & Smith and Wallis (2009) or Qian et al. (2019) and the references in these papers). Exemplarily, Claeskens et al. (2016) proved for the situation of forecasting one value of a time series that there is no guarentee that model averaging with random weights (not only smooth AIC-weights, but also other random weights) will be better than the model averaging estimator with fixed uniform weights. However, we observe that for large sample sizes the estimator after model selection and the model averaging estimator with smooth AIC-weights behave similar and outperform the model averaging estimator with uniform weights due to their asymptotic properties.

2.2.2. Models of more different shape. We will now consider the candidate set $\mathcal{S}_2$ in Eq. 2.9, which serves as an example of more different models and includes the log-linear, the Emax and the exponential model. The simulated mean squared errors of the three estimates of the $\mathrm{ED}_{0.4}$ are given in Table 3.

In contrast to Section 2.2.1 we observe only one scenario, where model averaging with uniform weights gives the smallest mean squared error (but in this case model averaging with smooth AIC-weights yields very similar

results). If the sample size increases model averaging with smooth AIC-weights and estimation after model selection yield a substantially smaller mean squared error. An explanation of this observation consists in the fact that for a candidate set containing models with a rather different shape model averaging with uniform weights produces a large bias. On the other hand model averaging with smooth AIC-weights and estimation after model selection adapt to the data and put more weight on the "true" model, in particular if the sample size is large. As estimation after model selection has a larger variance and the variance is decreasing with increasing sample size, the bias is dominating the mean squared error for large sample sizes and thus estimation in the model selected by the AIC is more efficient for large sample sizes.

The numerical study in Sections 2.2.1 and 2.2.2 can be summarized as follows. The results observed in the literature have to be partially relativized. If the candidate set is a subset of commonly used dose-response curves as in Table 1, the superiority of model averaging with uniform weights can only be observed for classes of "similar" competing models and a not too large signal to noise ratio. On the other hand if the dose-response models in the candidate set are of rather different structure or the sample size is large (leading to a small signal to noise ratio), model averaging with data dependent weights (such as smooth AIC-weights) or estimation after model selection may show a better performance. For these reasons we will investigate optimal/efficient designs for all three estimation methods in the following sections. We will demonstrate that a careful design of experiments can improve the accuracy of all three estimates substantially.

3. Asymptotic Properties and Optimal Design

In this section we will derive the asymptotic properties of model averaging estimates with fixed weights in the case where the competing models are not nested. The results can be used for (at least) two purposes. On the one hand they provide some understanding of the empirical findings in Section 2, where we observed, that for increasing sample size the mean squared error of model averaging estimates is dominated by its bias. On the other hand, we will use these results to develop an asymptotic representation of the mean squared error of the model averaging estimate, which can be used for the construction of optimal designs.

3.1. Model Averaging for Non-Nested Models Hjort and Claeskens, (2003) provide an asymptotic distribution of frequentist model averaging

estimates making use of local alternatives which require the true data generating process to lie inside a wide parametric model. All candidate models are sub-models of this wide model and the deviations in the parameters are restricted to be of order $n^{-1/2}$. Using this assumption results in convenient approximations for the mean squared error as variance and bias are both of order $O(1/n)$. However, in the discussion of this paper Raftery and Zheng (2003) pose the question if the framework of local alternatives is realistic. More importantly, frequentist model averaging is also often used for non-nested models (see for example Verrier et al., 2014). In this section we will develop asymptotic theory for model averaging estimation in non-nested models. In particular, we do not assume that the "true" model is among the candidate models used in the model averaging estimate.

As we will apply our results for the construction of efficient designs for model averaging estimation we use the common notation of this field. To be precise, let Y denote a response variable and let x denote a vector of explanatory variables defined on a given compact design space $\mathcal{X}$. Suppose that Y has a density $g(y \mid x)$ with respect to a dominating measure. For estimating a quantity of interest, say μ, from the distribution g we use r different parametric candidate models with densities

$$f_1(y \mid x, \theta_1), \ldots, f_r(y \mid x, \theta_r) \tag{3.1}$$

where θ_s denotes the parameter in the sth model, which varies in a compact parameter space, say $\Theta_s \subset \mathbb{R}^{p_s}$ $(s = 1, ..., r)$. Note, that in general we do not assume that the density g is contained in the set of candidate models in Eq. 3.1 and that the regression model (2.1) investigated in Section 2 is a special case of this general notation.

We assume that k different experimental conditions, say $x_1, \ldots, x_k$, can be chosen in a design space $\mathcal{X}$ and that at each experimental condition x_i one can observe n_i responses, say $y_{i1}, \ldots, y_{in_i}$ (thus the total sample size is $n = \sum_{i=1}^{k} n_i$), which are realizations of independent identically distributed random variables $Y_{i1}, \ldots, Y_{in_i}$ with density $g(\cdot \mid x_i)$. For example, if g coincides with f_s then the density of the random variables $Y_{i1}, \ldots, Y_{in_i}$ is given by $f_s(\cdot \mid x_i, \theta_s)$ $(i = 1, \ldots, k)$. To measure efficiency and to compare different experimental designs we will use asymptotic arguments and consider the case $\lim_{n\to\infty} \frac{n_i}{n} = \xi_i \in (0, 1)$ for $i = 1, \ldots, k$. As common in optimal design theory we collect this information in the form

$$\xi = \{x_1, \ldots, x_k; \xi_1, \ldots, \xi_k\}, \tag{3.2}$$

which is called approximate design in the following discussion (see, for example, Kiefer, 1974). For an approximate design ξ of the form Eq. 3.2 and total

sample size n a rounding procedure is applied to obtain integers n_i taken at each x_i $(i = 1, \ldots, k)$ from the not necessarily integer valued quantities $\xi_i n$ (see, for example Pukelsheim (2006), Chapter 12).

The asymptotic properties of the maximum likelihood estimate (calculated under the assumption that f_s is the correct density) is derived under certain assumptions of regularity (see the Assumptions (A1)-(A6) in Appendix). In particular, we assume that the functions f_s are twice continuously differentiable with respect to θ_s and that several expectations of derivatives of the log-densities exist. For a given approximate design ξ and a candidate density f_s we denote by

$$\mathrm{KL}(g : f_s \mid \theta_s, \xi) = \int g(y \mid x) \log\left(\frac{g(y \mid x)}{f_s(y \mid x, \theta_s)}\right) dy d\xi(x), \tag{3.3}$$

the Kullback-Leibler divergence between the models g and f_s and assume that

$$\theta_{s,g}^*(\xi) = \arg\min_{\theta_s \in \Theta_s} \mathrm{KL}(g : f_s \mid \theta_s, \xi) \tag{3.4}$$

is unique for each $s \in \{1, \ldots, r\}$. For notational simplicity we will omit the dependency of the minimum on the density g, whenever it is clear from the context and denote the minimizer by $\theta_s^*(\xi)$. We also assume that the matrices

$$A_s(\theta_s, \xi) = \sum_{i=1}^{k} \xi_i \, \mathbb{E}_{g(\cdot|x_i)}\Big(\frac{\partial^2 \log f_s(Y_{ij} \mid x_i, \theta_s)}{\partial \theta_s \partial \theta_s^\top}\Big), \tag{3.5}$$

$$B_{st}(\theta_s, \theta_t, \xi) = \sum_{i=1}^{k} \xi_i \, \mathbb{E}_{g(\cdot|x_i)}\Big(\frac{\partial \log f_s(Y_{ij} \mid x_i, \theta_s)}{\partial \theta_s}\Big(\frac{\partial \log f_t(Y_{ij} \mid x_i, \theta_t)}{\partial \theta_t}\Big)^\top\Big), \tag{3.6}$$

exist, where expectations are taken with respect to the true distribution $g(\cdot \mid x_i)$.

Under standard assumptions White (1982) shows the existence of a measurable maximum likelihood estimate $\hat{\theta}_{n,s}$ for all candidate models which is strongly consistent for the (unique) minimizer $\theta_s^*(\xi)$ in Eq. 3.4. Moreover, the estimate is also asymptotically normal distributed, that is

$$\sqrt{n}(\hat{\theta}_{n,s} - \theta_s^*(\xi)) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, A_s^{-1}(\theta_s^*(\xi)) B_{ss}(\theta_s^*(\xi), \theta_s^*(\xi)) A_s^{-1}(\theta_s^*(\xi))\right), \tag{3.7}$$

where we assume the existence of the inverse matrices, $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution and we use the notations

$$A_s(\theta_s^*(\xi)) = A_s(\theta_s^*(\xi), \xi) \;, \quad B_{st}(\theta_s^*(\xi), \theta_t^*(\xi)) = B_{st}(\theta_s^*(\xi), \theta_t^*(\xi), \xi) \tag{3.8}$$

$(s, t = 1, \dots r)$. The following result gives the asymptotic distribution of model averaging estimates of the form Eq. 2.6.

Theorem 3.1. *If Assumptions (A1) - (A7) in Appendix are satisfied, then the model averaging estimate (2.6) satisfies*

$$\sqrt{n}\Big(\hat{\mu}_{mav} - \sum_{s=1}^{r} w_s \mu_s(\theta_s^*(\xi))\Big) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma_w^2(\theta^*(\xi))\right), \tag{3.9}$$

where the asymptotic variance is given by

$$\sigma_w^2(\theta^*(\xi)) = \sum_{s,t=1}^{r} w_s w_t \Big(\frac{\partial \mu_s(\theta_s^*(\xi))}{\partial \theta_s}\Big)^\top A_s^{-1}(\theta_s^*(\xi)) B_{st}\left(\theta_s^*(\xi), \theta_t^*(\xi)\right) A_t^{-1}\left(\theta_t^*(\xi)\right) \frac{\partial \mu_t(\theta_t^*(\xi))}{\partial \theta_t}. \tag{3.10}$$

Theorem 3.1 shows, that the model averaging estimate is biased for the true target parameter μ_{true}, unless we have $\sum_{s=1}^{r} w_s \mu_s(\theta_s^*(\xi)) = \mu_{\text{true}}$. Hence we aim to minimize the asymptotic mean squared error of the model averaging estimate. Note, that the bias does not depend on the sample size, while the variance is of order $O(1/n)$.

3.2. Optimal Designs for Model Averaging of Non-Nested Models Alhorn et al. (2019) determined optimal designs for model averaging minimizing the asymptotic mean squared error of the estimate calculated in a class of nested models under local alternatives and demonstrated that optimal designs lead to substantially more precise model averaging estimates than commonly used designs in dose finding studies. With the results of Section 3.1 we can develop a more general concept of design of experiments for model averaging estimation, which is applicable for non-nested models.

To be precise, we consider the criterion

$$\Phi_{\text{mav}}(\xi, g, \mu_{\text{true}}) = \frac{1}{n}\sigma_w^2(\theta^*(\xi)) + \Big(\sum_{s=1}^{r} w_s \mu_s(\theta_s^*(\xi)) - \mu_{\text{true}}\Big)^2 \approx \text{MSE}(\hat{\mu}_{\text{mav}}), \tag{3.11}$$

where μ_{true} is the target parameter in the "true" model with density g and $\sigma_w^2(\theta^*(\xi))$ and $\theta_s^*(\xi)$ are defined in Eqs. 3.10 and 3.4, respectively. Note that this criterion depends on the "true" distribution via μ_{true} and the best approximating parameters $\theta_s^*(\xi) = \theta_{s,g}^*(\xi)$.

For estimating the target parameter μ via a model averaging estimate of the form (2.6) most precisely a "good" design ξ yields small values of the criterion function $\Phi_{\text{mav}}(\xi, g, \mu_{\text{true}})$. Therefore, for a given finite set of candidate models $f_1, \dots, f_r$ and weights $w_s, s = 1, \dots, r$, a design ξ^* is called

locally optimal design for model averaging estimation of the parameter μ, if it minimizes the function $\Phi_{\text{mav}}(\xi, g, \mu_{true})$ in Eq. 3.11 in the class of all approximate designs on $\mathcal{X}$. Here the term "locally" refers to the seminal paper of Chernoff (1953) on optimal designs for nonlinear regression models, because the optimality criterion still depends the unkown density $g(y \mid x)$.

A general approach to address this uncertainty problem is a Bayesian approach based on a class of models for the density g. To be precise, let $\mathcal{G}$ denote a finite set of potential densities and let π denote a probability distribution on $\mathcal{G}$, then we call a design *Bayesian optimal design for model averaging estimation of the parameter* μ if it minimizes the function

$$\Phi^{\pi}_{\text{mav}}(\xi, \mu_{true}) = \int_{\mathcal{G}} \Phi_{\text{mav}}(\xi, g, \mu_{true}) d\pi(g). \tag{3.12}$$

In general, the set $\mathcal{G}$ can be constructed independently of the set of candidate models. However, in the context of model averaging it is reasonable to construct a class of potential models $\mathcal{G}$ from the candidate set as follows. We denote the candidate set of models in Eq. 3.1 by $\mathcal{S}$. Each of these models depends on a unknown parameter θ_s and we denote by $\mathcal{F}_{f_s} \subset \Theta_s$ a set of possible parameter values for the model f_s. Now let π_2 denote a prior distribution on $\mathcal{S}$ and for each $f_s \in \mathcal{S}$ let $\pi_1(\cdot \mid f_s)$ denote a prior distribution on $\mathcal{F}_{f_s}$. Finally, we define $\mathcal{G} = \{(g, \theta) : g \in \mathcal{S}, \theta \in \mathcal{F}_g\}$ and a prior

$$d\pi(g, \theta) = d\pi_1(\theta \mid g)\ d\pi_2(g), \tag{3.13}$$

then the criterion (3.12) can be rewritten as

$$\Phi^{\pi}_{\text{mav}}(\xi, \mu_{\text{true}}) = \int_{\mathcal{S}} \int_{\mathcal{F}_g} \Phi_{\text{mav}}(\xi, g, \mu_{true}) d\pi_1(\theta \mid g)\ d\pi_2(g), \tag{3.14}$$

In the finite sample study of the following section the set $\mathcal{S}$ and the set $\mathcal{F}_g$ (for any $g \in \mathcal{S}$) are finite, which results in a finite set $\mathcal{G}$.

Locally and Bayesian optimal designs for model averaging estimation have to be calculated numerically in all cases of practical interest. We will state now a necessary condition for the optimality of a given design with respect to the criterion ϕ^{π}_{mav}. Note, that this criterion is not convex and therefore a sufficient condition cannot be derived. In the following discussion

we denote by $A_s^* = A_s(\theta_{s,g}^*(\xi^*), \xi^*)$ and $B_{st}^* = B_{st}(\theta_{s,g}^*(\xi^*), \theta_{t,g}^*(\xi^*), \xi^*)$ the matrices defined in Eqs. 3.5 and 3.6, respectively, evaluated in ξ^* and $\theta_{s,g}(\xi^*)$.

Theorem 3.2. *If a design ξ^* is Bayesian optimal for model averaging estimation of the parameter μ with respect to the prior π, then*

$$d_\pi(x, \xi^*) = \tag{3.15}$$
$$\int_{\mathcal{G}} \frac{1}{n}\sigma_g'(\xi^*, x) + 2\Big(\sum_{s=1}^r w_s \mu_s(\theta_{s,g}^*(\xi^*)) - \mu_{true}\Big)\sum_{s=1}^r w_s \Big(\frac{\partial \mu_s(\theta_{s,g}^*(\xi^*))}{\partial \theta_s}\Big)^\top \theta_{s,g}'(\xi^*, x) d\pi(g) \leq 0$$

holds for all $x \in \mathcal{X}$, where the derivatives $\theta_{s,g}'(\xi^*, x)$ and $\sigma_g'(\xi^*, x)$ are given by

$$\begin{aligned}
\theta_{s,g}'(\xi^*, x) &= -\Big(\int\int g(y \mid t)\frac{\partial^2}{\partial \theta_s \partial \theta_s^\top} \log f_s(y \mid t, \theta_{s,g}^*(\xi^*)) dy d\xi^*(t)\Big)^{-1} \cdot \\
&\qquad \int g(y \mid x)\frac{\partial}{\partial \theta_s} \log f_s(y \mid x, \theta_{s,g}^*(\xi^*)) dy \qquad (3.16)\\
\sigma_g'(\xi^*, x) &= \sum_{s,t} w_s w_t \cdot \Big[\Big(\frac{\partial^2 \mu_s(\theta_{s,g}^*(\xi^*))}{\partial \theta_s \partial \theta_s^\top}\theta_{s,g}'(\xi^*, x)\Big)^\top (A_s^*)^{-1} B_{st}^* (A_t^*)^{-1}\frac{\partial \mu_t(\theta_{t,g}^*(\xi^*))}{\partial \theta_t} \qquad (3.17)\\
&\quad -\Big(\frac{\partial \mu_s(\theta_{s,g}^*(\xi^*))}{\partial \theta_s}\Big)^\top \big((A_s^*)^{-1} h_{s,g}'(\xi^*, x)(A_s^*)^{-1}\big) B_{st}^* (A_t^*)^{-1}\frac{\partial \mu_t(\theta_{t,g}^*(\xi^*))}{\partial \theta_t}\\
&\quad +\Big(\frac{\partial \mu_s(\theta_{s,g}^*(\xi^*))}{\partial \theta_s}\Big)^\top (A_s^*)^{-1} h_{st,g}'(\xi^*, x)(A_t^*)^{-1}\frac{\partial \mu_t(\theta_{t,g}^*(\xi^*))}{\partial \theta_t}\\
&\quad -\Big(\frac{\partial \mu_s(\theta_{s,g}^*(\xi^*))}{\partial \theta_s}\Big)^\top (A_s^*)^{-1} B_{st}^* \left((A_t^*)^{-1} h_{t,g}'(\xi^*, x)(A_t^*)^{-1}\right)\frac{\partial \mu_t(\theta_{t,g}^*(\xi^*))}{\partial \theta_t}\\
&\quad +\Big(\frac{\partial \mu_s(\theta_{s,g}^*(\xi^*))}{\partial \theta_s}\Big)^\top (A_s^*)^{-1} B_{st}^* (A_t^*)^{-1}\frac{\partial^2 \mu_t(\theta_{t,g}^*(\xi^*))}{\partial \theta_t \partial \theta_t^\top}\theta_{t,g}'(\xi^*, x)\Big],
\end{aligned}$$

respectively. Here the matrices $h_{st,g}'(\xi^*, x)$ and $h_{s,g}'(\xi^*, x)$ are given by

$$\begin{aligned}
h_{st,g}'(\xi^*, x) &= \int\int g(y \mid u)\left[\frac{\partial^2 \log f_s(y \mid u, \theta_{s,g}^*(\xi^*))}{\partial \theta_s \partial \theta_s^\top}\theta_{s,g}'(\xi^*, x)\left(\frac{\partial \log f_t(y \mid u, \theta_{t,g}^*(\xi^*))}{\partial \theta_t}\right)^\top\right.\\
&\quad \left. + \frac{\partial \log f_s(y \mid u, \theta_{s,g}^*(\xi^*))}{\partial \theta_s}\left(\theta_{t,g}'(\xi^*, x)\right)^\top\left(\frac{\partial^2 \log f_t(y \mid u, \theta_{t,g}^*(\xi^*))}{\partial \theta_t \partial \theta_t^\top}\right)^\top\right] dy d\xi^*(u)\\
&\quad + B_{st}(\theta_{s,g}^*(\xi^*), \theta_{t,g}^*(\xi^*), \xi_x) - B_{st}(\theta_{s,g}^*(\xi^*), \theta_{t,g}^*(\xi^*), \xi^*) \qquad (3.18)\\
h_{s,g}'(\xi^*, x) &= \int\int g(y \mid u) D_s(\theta_{s,g}^*(\xi^*))(I_{p_s} \otimes \theta_{s,g}'(\xi^*, x)) dy d\xi^*(u)\\
&\quad + A_s(\theta_{s,g}^*(\xi^*), \xi_x) - A_s(\theta_{s,g}^*(\xi^*), \xi^*) \qquad (3.19)
\end{aligned}$$

where the matrix

$$D_s(\theta_s) = \begin{pmatrix} \frac{\partial^3 \log f_s(y|x,\theta_s)}{\partial\theta_{s,1}\partial\theta_{s,1}\partial\theta_{s,1}} & \cdots & \frac{\partial^3 \log f_s(y|x,\theta_s)}{\partial\theta_{s,1}\partial\theta_{s,1}\partial\theta_{s,p_s}} & \cdots & \frac{\partial^3 \log f_s(y|x,\theta_s)}{\partial\theta_{s,1}\partial\theta_{s,p_s}\partial\theta_{s,1}} & \cdots & \frac{\partial^3 \log f_s(y|x,\theta_s)}{\partial\theta_{s,1}\partial\theta_{s,p_s}\partial\theta_{s,p_s}} \\ \frac{\partial^3 \log f_s(y|x,\theta_s)}{\partial\theta_{s,2}\partial\theta_{s,1}\partial\theta_{s,1}} & \cdots & \frac{\partial^3 \log f_s(y|x,\theta_s)}{\partial\theta_{s,2}\partial\theta_{s,1}\partial\theta_{s,p_s}} & \cdots & \frac{\partial^3 \log f_s(y|x,\theta_s)}{\partial\theta_{s,2}\partial\theta_{s,p_s}\partial\theta_{s,1}} & \cdots & \frac{\partial^3 \log f_s(y|x,\theta_s)}{\partial\theta_{s,2}\partial\theta_{s,p_s}\partial\theta_{s,p_s}} \\ \vdots & & \vdots & & \vdots & & \vdots \\ \frac{\partial^3 \log f_s(y|x,\theta_s)}{\partial\theta_{s,p_s}\partial\theta_{s,1}\partial\theta_{s,1}} & \cdots & \frac{\partial^3 \log f_s(y|x,\theta_s)}{\partial\theta_{s,p_s}\partial\theta_{s,1}\partial\theta_{s,p_s}} & \cdots & \frac{\partial^3 \log f_s(y|x,\theta_s)}{\partial\theta_{s,p_s}\partial\theta_{s,p_s}\partial\theta_{s,1}} & \cdots & \frac{\partial^3 \log f_s(y|x,\theta_s)}{\partial\theta_{s,p_s}\partial\theta_{s,p_s}\partial\theta_{s,p_s}} \end{pmatrix}$$

contains the third derivatives of the log-likelihood with respect to the parameters $\theta_s = (\theta_{s,1}, \ldots, \theta_{s,p_s})^\top$. *Moreover, there is equality in Eq. 3.15 for all support points of the optimal design.*

Example 3.1. *We illustrate the application of Theorem 3.2 for regression models of the from Eq. 2.1 with centred normal distributed errors. As regression functions we use the log-linear and the Emax model and their parameter specifications given in Table 1. Then, the locally optimal designs for estimation of the* $ED_{0.4}$ *in the log-linear model* f_1 *and in the Emax model* f_2 *are given by*

$$\xi_2 = \{0, 4.051, 150;\ 0.339, 0.5, 0.161\}, \tag{3.20}$$

and $\{0, 18.75, 150;\ 0.25, 0.5, 0.25\}$, *respectively see Dette et al. (2010). For sample size* $n = 100$ *we determine a Bayesian optimal design for model averaging estimation of the* $ED_{0.4}$ *(with uniform weights) with respect to the criterion* (3.12). *The set of possible models is given by* $\mathcal{G} = \{f_1, f_2\}$ *with parameters specified in Table 1, and we choose a uniform prior on this set. The optimal design has been calculated numerically using the COBYLA algorithm (see* (Powell, 1994)*) and is given by*

$$\xi_{12}^* = \{0, 13.026, 150;\ 0.281, 0.498, 0.220\}. \tag{3.21}$$

The necessary condition of Theorem 3.2 is satisfied as illustrated in Fig. 2. Note that the design ξ_{12}^* *can be considered as a compromise between the locally optimal designs for the individual models and that* ξ_{12}^* *would not be optimal if the inequality was not satisfied.*

We conclude noting that the optimality criteria proposed in this section have been derived for model averaging estimates with fixed weights. The asymptotic theory presented here cannot be easily adapted to estimates using data-dependent (random) weights (as considered in Section 2), because it is difficult to get an explicit expression for the asymptotic distribution, which

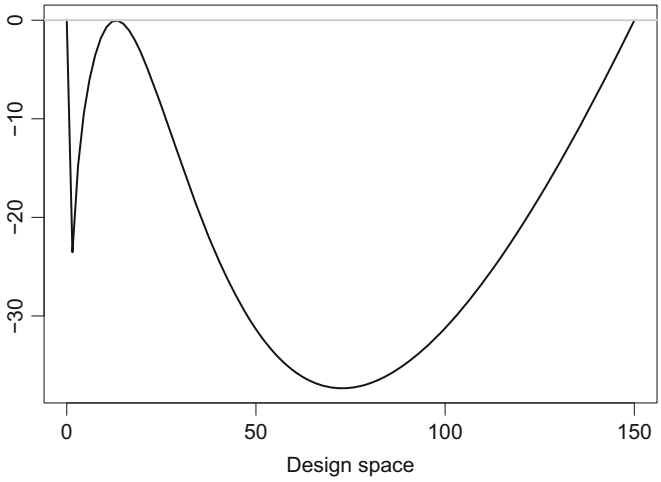

Figure 2: Necessary condition of Theorem 3.2 for the optimal design (3.21)

is not normal in general. Nevertheless, we will demonstrate in the following section that designs minimizing the mean squared error of model averaging estimates with fixed weights will also yield a substantial improvement in model averaging estimation with smooth AIC-weights and in estimation after model selection.

4 Bayesian Optimal Designs For Model Averaging

We will demonstrate by means of a simulation study that the performance of all considered estimates can be improved substantially by the choice of an appropriate design. For this purpose we consider the same situation as in Section 2, that is regression models of the form (2.1) with centred normal distributed errors. We also consider the two different candidate sets $\mathcal{S}_1$ and $\mathcal{S}_2$ defined in Eq. 2.8 (log-linear, Emax and quadratic model) and Eq. 2.9 (log-linear, Emax and exponential model), respectively.

Using the criterion introduced in Section 3 we now determine a Bayesian optimal design for model averaging estimation of the $\mathrm{ED}_{0.4}$ with uniform weights from $n = 100$ observations. Note that we use the sample size $n = 100$ since this is a common available sample size in the context of dose finding studies. We require a prior distribution for the unknown density g, and we use a distribution of the form Eq. 3.13 for this purpose. To be precise, let $f_s(y \mid x, \theta_s)$ denote the density of a normal distribution with mean $\eta_s(x, \vartheta_s)$ and variance $\sigma_s^2 = 0.1$ $(s = 1, \ldots, r)$, where the mean functions are given in Table 1. As the criterion (3.14) does not depend on the intercept ϑ_{s1}, these

are not varied and taken from Table 1. For each of the other parameters we use three different values: the values specified in Table 1 and a 10% larger and smaller value of this parameter.

$$\begin{aligned}
\mathcal{F}_{f_1} &= \{(0, \vartheta_{12}, \vartheta_{13}) : \vartheta_{12} = 0.0797 \pm 10\%, \vartheta_{13} = 1 \pm 10\%\}, \\
\mathcal{F}_{f_2} &= \{(0, \vartheta_{22}, \vartheta_{23}) : \vartheta_{22} = 0.467 \pm 10\%, \vartheta_{23} = 25 \pm 10\%\}, \\
\mathcal{F}_{f_3} &= \{(-0.08265, \vartheta_{32}, \vartheta_{33}) : \vartheta_{32} = 0.08265 \pm 10\%, \vartheta_{33} = 85 \pm 10\%\}, \\
\mathcal{F}_{f_4} &= \{(0, \vartheta_{42}, \vartheta_{43}) : \vartheta_{42} = 0.00533 \pm 10\%, \vartheta_{43} = -0.00002 \pm 10\%\}.
\end{aligned} \tag{4.1}$$

4.1. Models of similar shape We will first consider the candidate set $\mathcal{S}_1 = \{f_1, f_2, f_4\}$ consisting of the log-linear, the Emax and the quadratic model. For the definition of the prior distribution (3.13) in the criterion (3.14) we consider a uniform distribution π_2 on the set $\mathcal{S}_1$ and a uniform prior $\pi_1(\cdot \mid f_s)$ on each set $\mathcal{F}_{f_s}$ in Eq. 4.1 ($s = 1, 2, 4$). The Bayesian optimal design for model averaging estimation of the $\mathrm{ED}_{0.4}$ minimizing the criterion (3.14) is given by

$$\xi^*_{\mathcal{S}_1} = \{0, 18.310, 67.102, 150;\ 0.205, 0.290, 0.281, 0.224\}\,. \tag{4.2}$$

We will compare this design with the design

$$\xi_1 = \{0, 10, 25, 50, 100, 150;\ 1/6, 1/6, 1/6, 1/6, 1/6, 1/6\}\,, \tag{4.3}$$

proposed in Pinheiro et al. (2006) for a a similar setting (this design has also been used in Section 2) and the locally optimal design for the estimation of the $\mathrm{ED}_{0.4}$ in the log-linear model given by Eq. 3.20.

Results for the locally optimal designs for estimation of the $\mathrm{ED}_{0.4}$ in the Emax and exponential model are similar and omitted for the sake of brevity. We use the same setup as in Section 2.

The corresponding results are given in Table 4, where we use the models f_1, f_2 and f_4 from Table 1 to generate the data. The different columns represent the different estimation methods (left column: model averaging with uniform weights; middle column: smooth AIC-weights, right column: model selection). The numbers printed in boldface indicate the minimal mean squared error for each estimation method obtained from the different experimental designs. Compared to the designs ξ_1 and ξ_2 the Bayesian optimal design $\xi^*_{\mathcal{S}_1}$ for model averaging with uniform weights improves the efficiency of all estimation techniques. For example, when data is generated using the log-linear model f_1 the mean squared error of the model averaging estimate with uniform weights is reduced by 20.5% and 4.2%, when the optimal design is used instead of the designs ξ_1 or ξ_2, respectively. This

Table 4: Simulated mean squared errors of different estimates of the $ED_{0.4}$ for different experimental designs

Model	Design	Uniform weights	Smooth AIC-weights	Model selection
	Eq. 4.2	**177.472**	**165.981**	**173.548**
f_1	Eq. 4.3	223.291	218.990	285.062
	Eq. 3.20	185.251	184.77	340.698
	Eq. 4.2	**142.085**	**153.745**	**170.059**
f_2	Eq. 4.3	189.785	203.796	251.836
	Eq. 3.20	501.814	501.394	1162.654
	Eq. 4.2	**160.039**	**195.116**	**299.365**
f_4	Eq. 4.3	190.662	244.558	391.443
	Eq. 3.20	404.716	427.548	1396.051

The set of candidate models is $\mathcal{S}_1 = \{f_1, f_2, f_4\}$. Left column: model averaging estimate with uniform weights; middle column: model averaging estimate with smooth AIC-weights; right column: estimate after model selection

improvement is remarkable as the design ξ_2 is locally optimal for estimating the $ED_{0.4}$ in the model f_1 and data is generated from this model. In other cases the improvement is even more visible. For example, if data is generated by the model f_2 the improvement in model averaging estimation with uniform weights is 25.1% and 71.7% compared to the designs ξ_1 and ξ_2 defined in Eqs. 4.3 and 3.20. Moreover, although the designs are constructed for model averaging with uniform weights they also yield substantially more accurate model averaging estimates with smooth AIC-weights and a more precise estimate after model selection. For example, if the data is generated from model f_1 the mean squared error is reduced by 24.2% and by 10.2% for estimation with smooth AIC-weights and by 39.1% and 49.1% for estimation after model selection, respectively. Similar results can be observed for the models f_2 and f_4.

Summarizing, our numerical results show that the Bayesian optimal design for model averaging estimation of the $ED_{0.4}$ yields a substantial improvement of the mean squared error of the model averaging estimate with uniform weights (4.2%-71.7%), smooth AIC-weights (10.2%-69.3%) and the estimate after model selection (23.5%-85.4%) for all three models under consideration.

4.2. Models of Different Shape We will now consider the second candidate set $\mathcal{S}_2$ consisting of the log-linear (f_1) the Emax (f_2) and the exponential model (f_3). For the definition of the prior distribution (3.13) in the criterion (3.14) we use a uniform distribution π_2 on the set $\mathcal{S}_2$ and a uniform prior $\pi_1(\cdot \mid f_s)$ on each set $\mathcal{F}_{f_s}$ $(s = 1, 2, 3)$ in Eq. 4.1. For this choice the Bayesian

optimal design for model averaging estimation of the $\mathrm{ED}_{0.4}$ is given by

$$\xi^*_{\mathcal{S}_2} = \{0, 10.025, 77.746, 84.556, 150;\ 0.192, 0.212, 0.198, 0.189, 0.208\}\,, \quad (4.4)$$

and has (in comparison to the design $\xi^*_{\mathcal{S}_1}$ in Section 4.1) five instead of four support points.

The simulated mean squared errors of the three estimates under different designs are given in Table 5. We observe again that compared to the designs ξ_1 and ξ_2 in Eqs. 4.3 and 3.20 the Bayesian optimal design $\xi^*_{\mathcal{S}_2}$ improves most estimation techniques substantially. However, if model averaging with uniform weights is used and data is generated by model f_2 or f_3, the mean squared error of the model averaging estimate from the optimal design is 5.4% and 4.5% larger than the mean squared error obtained by the design ξ_1, respectively. For model averaging with smooth AIC-weights and data being generated from model f_2 this difference is 5.9%. Overall, the reported results demonstrate a substantial improvement in efficiency by usage of the Bayesian optimal design independently of the estimation method. If the Bayesian optimal design is used, estimation after model selection yields the smallest mean squared error if the data is generated from a model of the candidate set $\mathcal{S}_2$.

Summarizing, our numerical results show that compared to the designs ξ_1 and ξ_2 the design $\xi^*_{\mathcal{S}_2}$ reduces the mean squared error of model averaging estimates with uniform weights up to 50.3%. Furthermore, for smooth AIC-weights and estimation after model selection the reduction can be even larger and is up to 70.5% and 85.3%, respectively.

Table 5: Simulated mean squared errors of different estimates of the $\mathrm{ED}_{0.4}$ for different experimental designs

Model	Design	Uniform weights	Smooth AIC-weights	Model selection
	Eq. 4.4	**654.914**	**279.257**	**274.016**
f_1	Eq. 4.3	712.404	340.254	353.707
	Eq. 3.20	770.705	410.715	413.676
	Eq. 4.4	546.105	283.757	**250.719**
f_2	Eq. 4.3	**517.963**	**267.967**	286.272
	Eq. 3.20	1098.323	962.257	1701.569
	Eq. 4.4	910.372	**742.507**	**699.612**
f_3	Eq. 4.3	**871.362**	766.140	802.763
	Eq. 3.20	1505.693	1774.895	2592.261

The set of candidate models is $\mathcal{S}_2 = \{f_1, f_2, f_3\}$. Left column: model averaging estimate with uniform weights; middle column: model averaging estimate with smooth AIC-weights; right column: estimate after model selection

5 Robustness of the Designs

The designs determined in Sections 4.1 and 4.2 are Bayesian optimal for estimating the $ED_{0.4}$ under the assumption that the true data generating model is part of the set of candidate models $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively.

In the following we will analyse the behavior of these designs if these assumptions are not completely satisfied. More precisely, in Section 5.1 we investigate the performance of the different estimators using the designs (4.2) and (4.4) if the underlying true model is not part of the candidate sets. In Section 5.2 we consider the performance of these designs if they are used to estimate not only the $ED_{0.4}$, but also the $ED_{0.5}$ and $ED_{0.8}$. In this context, we also derive multi-objective designs which are recommended if more than one parameter of interest has to be estimated.

5.1. Robustness with Respect to Data Generating Model In this section we analyse the performance of the designs determined in Sections 4.1 and 4.2, if the true data generating model is not among the candidate models. More precisely, we consider the same setup as in Section 4.1 with candidate set $\mathcal{S}_1 = \{f_1, f_2, f_4\}$ and the corresponding design in Eq. 4.2 where we use the model f_3 to generate the data, and the setup as in Section 4.2 with candidate set $\mathcal{S}_2 = \{f_1, f_2, f_3\}$ and the corresponding design in Eq. 4.4 where we use the model f_4 to generate the data, respectively.

The corresponding results are presented in Tables 6 and 7, respectively. Compared to the designs ξ_1 (see Eq. 4.3) and ξ_2 (see Eq. 3.20) the Bayesian optimal designs still improve substantially the efficiency of all estimation techniques, although the true data generating models are not contained the candidate sets used in the definition of the corresponding optimality criterion.

In the setup of Section 4.1 where the candidate set $\mathcal{S}_1$ with corresponding optimal design is used (see Table 6), the improvement is less pronounced for model averaging with uniform weights (4.6% and 66.8% compared to the designs ξ_1 and ξ_2 in Eqs. 4.3 and 3.20, respectively) than for smooth AIC-weights (10.5% and 77.5%) and estimation after model selection (16.9% and 85.2%). Considering the setup of Section 4.2 where the candidate set $\mathcal{S}_2$ with corresponding optimal design is used (see Table 7), the improvement of the model averaging methods with uniform weights (23.4% and 69.4% compared to the designs ξ_1 and ξ_2 in Eqs. 4.3 and 3.20, respectively) and estimation after model selection (17.3% and 81.2%) are most obvious.

Moreover, we observe that the model averaging estimator with uniform weights is outperformed by the model averaging estimator with smooth AIC weights and the estimate after model selection in the setup of Section 4.1 (see Table 6), whereas it is the other way around in the setup of Section 4.2

Table 6: Simulated mean squared errors of different estimates of the $ED_{0.4}$ for different experimental designs

Model	Design	Uniform weights	Smooth AIC-weights	Model selection
	Eq. 4.2	**1058.655**	**766.408**	**606.752**
f_3	Eq. 4.3	1109.622	856.484	729.912
	Eq. 3.20	3184.11	3413.566	4102.964

The set of candidate models is $\mathcal{S}_1 = \{f_1, f_2, f_4\}$. Left column: model averaging estimate with uniform weights; middle column: model averaging estimate with smooth AIC-weights; right column: estimate after model selection

(see Table 7), where the model averaging estimator with uniform weights performs best. The good performance of model averaging estimates with uniform weights can also be observed in other settings where all candidate models misspecify the true data generating model. As stated in Section 2.2.1, several theoretical and heuristical results about this phenomenon were deduced especially in the context of forecasting of time series and we again refer to Qian et al. (2019) for a good review on that issue.

Summarizing, the Bayesian optimal designs still improve the accurarcy of all estimation techniques even if the true data generating model is not among the candidate models.

5.2. Robustness with respect to parameter of interest In the previous sections we assumed that there is one target parameter μ and the considered Bayesian optimal designs were supposed to improve the performance of the three estimation methods for μ. In the following, we briefly indicate how this methodology can be further extended to adress the problem of estimating several target parameters, say $\mu^{(1)}, \ldots, \mu^{(L)}$. For this purpose we follow the idea of Kao et al. (2009) and define a multi-objective criterion by an average of the criteria resulting from the individual target parameters.

Table 7: Simulated mean squared errors of different estimates of the $ED_{0.4}$ for different experimental designs

Model	Design	Uniform weights	Smooth AIC-weights	Model selection
	Eq. 4.4	**159.899**	**278.409**	**347.187**
f_4	Eq. 4.3	208.628	298.315	419.651
	Eq. 3.20	522.652	610.198	1907.066

The set of candidate models is $\mathcal{S}_2 = \{f_1, f_2, f_3\}$. Left column: model averaging estimate with uniform weights; middle column: model averaging estimate with smooth AIC-weights; right column: estimate after model selection

More precisely, we consider a similar setup as in Section 3.2, that is, $\mathcal{G}$ is a finite set of potential densities and π is a probability distribution on $\mathcal{G}$. We call a design *multi-objective Bayesian optimal design for model averaging estimation of the parameters* $\mu^{(1)}, \ldots, \mu^{(L)}$ if it minimizes the function

$$\bar{\Phi}^{\pi}_{\text{mav}}(\xi) = \frac{1}{L} \sum_{\ell=1}^{L} \int_{\mathcal{G}} \Phi_{\text{mav}}(\xi, g, \mu^{(\ell)}) d\pi(g), \tag{5.1}$$

where $\Phi_{\text{mav}}(\xi, g, \mu^{(\ell)})$ denotes the Bayesian optimality criterion defined in Eq. 3.12 depending on the individual target parameter $\mu^{(\ell)}$ $(\ell = 1, \ldots, L)$.

We now demonstrate by means of a simulation study that the designs based on the extended criterion defined in Eq. 5.1 can be useful to improve the performance of all considered estimation methods if several parameters are of interest. For the sake of brevity, we concentrate on the situation of Section 4.1, where the candidate set is given by $\mathcal{S}_1 = \{f_1, f_2, f_4\}$ (log-linear, Emax and quadratic model, cf. Table 1) with the corresponding uniform prior distribution given by a uniform distribution π_2 on the set $\mathcal{S}_1$ and by a uniform prior $\pi_1(\cdot \mid f_s)$ on each set $\mathcal{F}_{f_s}$ in Eq. 4.1 $(s = 1, 2, 4)$. Results for the setup used in Section 4.2 are similar and omitted for the sake of brevity. We will consider the problem of estimating the three target parameters $\mu^{(1)} = \text{ED}_{0.4}$, $\mu^{(2)} = \text{ED}_{0.5}$ and $\mu^{(3)} = \text{ED}_{0.8}$ as defined in Eq. 2.4 using the designs given in Eqs. 4.2, 4.3, and 3.20 on the one hand. On the other hand we will use the multi-objective Bayesian optimal design for model averaging estimation of the $\text{ED}_{0.4}$, $\text{ED}_{0.5}$ and $\text{ED}_{0.8}$ minimizing the criterion (5.1) which is given by

$$\bar{\xi}^{*}_{\mathcal{S}_1} = \{0, 15.437, 60.887, 150;\ 0.214, 0.301, 0.242, 0.242\}\,. \tag{5.2}$$

The simulated averages of the mean squared errors

$$\frac{1}{3} \sum_{\ell=1}^{3} \mathbb{E}\big[(\hat{\mu}^{(\ell)} - \mu^{(\ell)})^2\big]$$

of the estimates for the three target parameters $\text{ED}_{0.4}$, $\text{ED}_{0.5}$ and $\text{ED}_{0.8}$ under the different designs and different estimation methods are given in Table 8. Again, we use the same simulation setup as in Section 4.1. We observe that compared to the designs ξ_1 and ξ_2 in Eqs. 4.3 and 3.20 the multi-objective Bayesian optimal design $\bar{\xi}^{*}_{\mathcal{S}_1}$ improves most estimation techniques. However, if model averaging with smooth AIC weights is used and data is

Table 8: The average of the simulated mean squared errors of the different estimates of the $ED_{0.4}$, $ED_{0.5}$, $ED_{0.8}$ for different experimental designs

Model	Design	Uniform weights	Smooth AIC-weights	Model selection
	Eq. 5.2	**278.958**	330.105	**355.499**
f_1	Eq. 4.2	298.454	354.495	384.584
	Eq. 4.3	334.452	386.926	434.685
	Eq. 3.20	304.829	**308.245**	578.353
	Eq. 5.2	**272.069**	**299.87**	**327.625**
f_2	Eq. 4.2	297.191	325.94	341.329
	Eq. 4.3	309.732	339.03	365.93
	Eq. 3.20	804.824	821.560	1522.018
	Eq. 5.2	281.083	335.163	465.125
f_4	Eq. 4.2	**277.695**	**327.043**	**440.443**
	Eq. 4.3	282.828	338.638	482.282
	Eq. 3.20	644.253	731.287	1674.672

The set of candidate models is $\mathcal{S}_1 = \{f_1, f_2, f_4\}$. Left column: model averaging estimate with uniform weights; middle column: model averaging estimate with smooth AIC-weights; right column: estimate after model selection

generated by the log-linear model f_1, the average of the mean squared errors is 7.09% larger than the average of the mean squared errors obtained by design ξ_2 which is locally optimal for estimation of the $ED_{0.4}$ in the log-linear model. Moreover, we observe that the Bayesian optimal design for model averaging estimation of $ED_{0.4}$ in Eq. 4.2 yields similar results as the multi-objective Bayesian design in Eq. 5.1. In the case where data is generated by the quadratic model f_4 the design $\xi^*_{\mathcal{S}_1}$ even improves the mean squared error of all three estimation techniques compared to the design $\bar{\xi}^*_{\mathcal{S}_1}$. Consequently, the design $\xi^*_{\mathcal{S}_1}$, which is supposed to result in a precise estimation of the $ED_{0.4}$, is robust with respect to variations of the target parameter and can also be used for efficient estimation of other ED_p values.

Nevertheless, the criterion defined in Eq. 5.1 can be useful if the focus is widened to the estimation of more different parameters, for instance the estimation of the $ED_{0.4}$ and the prediction of an effect at a prespecified dose level.

6 Conclusions

In this paper we derived the asymptotic distribution of the frequentist model averaging estimate with fixed weights from a class of not necessarily nested models.

We use these results to determine Bayesian optimal designs for model averaging, which can improve the estimation accuracy of the estimate substantially. Although these designs are constructed for model averaging with fixed weights, they also yield a substantial improvement of accuracy for model averaging with data dependent weights and for estimation after model selection.

We also demonstrate that the superiority of model averaging against estimation after model selection in the context of dose finding studies depends sensitively on the class of competing models, which is used in the model averaging procedure. If the competing models are similar (which means that a given model from the class can be well approximated by all other models) and the signal to noise ratio is large, then model averaging should be preferred. Otherwise, we observe advantages for estimation after model selection, in particular, if the signal to noise ratio is small.

Although, the new designs show a very good performance for estimation after model selection and for model averaging with data dependent weights, it is of interest to develop optimal designs, which address the specific issues of data dependent weights in the estimates. This is a very challenging problem for future research as there is no simple expression of the asymptotic mean squared error of these estimates. A first approach to solve this problem is an adaptive one and a further interesting and very challenging question of future research is to improve the accuracy of adaptive designs.

Moreover, in the present paper we only briefly discuss the situation where the true data generating model is not among the candidate models. In this situation different estimation strategies might be suitable, such as the use of data-dependent weights for model combining which directly take the minimization of the mean squared error into account (see Qian et al. (2019) & Zhang et al. (2016) and Wang et al. (2009) among many others) or adaptive approaches, which work both for the parametric candidate and for nonparametric models (see Yang 2001, 2003). Consequently, another interesting problem for future research will be the construction of optimal designs for these estimators.

A further extremely challenging topic of future research is the construction of designs for different estimation techniques in big data analysis (such as convolutional neural networks or random forests). In such applications the focus is on (sub-)sampling and the construction of design strategies for a fixed model and a given estimation technique is just at the beginning of its development (see Ma et al. (2015) & Wang et al. (2019b) or Wang 2019a). An extension of these (sub-)sampling techniques to the case of multiple models and different estimation techniques is of particular practical importance.

Acknowledgments. This work has also been supported in part by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823, Teilprojekt C2, T1) of the German Research Foundation (DFG). The authors would like to thank the referee for his constructive comments on an earlier version of this paper.

Funding. Open Access funding enabled and organized by Projekt DEAL.

References

ALHORN, K., SCHORNING, K. and DETTE, H. (2019). Optimal designs for frequentist model averaging. *Biometrika* **106**, 3, 665–682.

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

BATES, J. M. and GRANGER, C. W. J. (1969). The combination of forecasts. *OR* **20**, 4, 451–468.

BOX, G. E. P. and HILL, W. J. (1967). Discrimination among mechanistic models. *Technometrics* **9**, 1, 57–71.

BRETZ, F., HSU, J. and PINHEIRO, J. (2008-07). Dose finding – a challenge in statistics. *Biom. J.* **50**, 4, 480–504.

BUATOIS, S., UECKERT, S., FREY, N., RETOUT, S. and MENTRÉ, F. (2018). Comparison of model averaging and model selection in dose finding trials analyzed by nonlinear mixed effect models. *The AAPS Journal* **20**, 56.

BUCKLAND, S. T., BURNHAM, K. P. and AUGUSTIN, N. H. (1997). Model selection: An integral part of inference. *Biometrics* **53**, 2, 603–618.

BURNHAM, K. P. and ANDERSON, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*, 2nd edn. Springer, New York.

CHEN, L., WAN, A. T. K., TSO, G. and ZHANG, X. (2018). A model averaging approach for the ordered probit and nested logit models with applications. *Journal of Applied Statistics* **45**, 16, 3012–3052.

CHERNOFF, H. (1953). Locally optimal designs for estimating parameters. *Annals of Mathematical Statistics* **24**, 586–602.

CLAESKENS, G. and HJORT, N. L. (2008). *Model selection and model averaging. Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

CLAESKENS, G., MAGNUS, J. R., VASNEV, A. and WANG, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting* **32**, 3, 754–762.

DETTE, H. (1990). A generalization of D- and D_1-optimal designs in polynomial regression. *The Annals of Statistic* **18**, 1784–1805.

DETTE, H., KISS, C., BEVANDA, M. and BRETZ, F. (2010). Optimal designs for the emax, log-linear and exponential models. *Biometrika* **97**, 2, 513–518.

DETTE, H., MELAS, V. B. and GUCHENKO, R. (2015). Bayesian t-optimal discriminating designs. *The Annals of Statistics* **43**, 5, 1959–1985.

HANSEN, B. E. (2007-07). Least squares model averaging. *Econometrica* **75**, 4, 1175–1189.

HJORT, N. L. and CLAESKENS, G. (2003-12). Frequentist Model Average Estimators. *Journal of the American Statistical Association* **98**, 464, 879–899.

HOETING, J. A., MADIGAN, D., RAFTERY, A. E. and VOLINSKY, C. T. (1999). Bayesian model averaging: a tutorial (with comments by m. clyde, david draper and e. i. george, and a rejoinder by the authors). *Statist. Sci.* **14**, 4, 382–417.

HONG, H. and PRESTON, B. (2012). Bayesian averaging, prediction and nonnested model selection. *Journal of Econometrics* **167**, 2, 358–369. Fourth Symposium on Econometric Theory and Applications (SETA).

KAO, M-H, MANDAL, A., LAZAR, N. and STUFKEN, J. (2009). Multi-objective optimal experimental designs for event-related fmri studies. *Neuroimage* **44**, 3, 849–856.

KAPETANIOS, G., LABHARD, V. and PRICE, S. (2008). Forecasting using bayesian and information-theoretic model averaging. *Journal of Business & Economic Statistics* **26**, 1, 33–41.

KIEFER, J. (1974-09). General Equivalence Theory for Optimum Designs (Approximate Theory). *The Annals of Statistics* **2**, 5, 849–879.

KONISHI, S. and KITAGAWA, G. (2008). *Information criteria and statistical modeling*. Wiley, New York.

LIANG, H., ZOU, G., WAN, A. T. K. and ZHANG, X. (2011-09). Optimal Weight Choice for Frequentist Model Average Estimators. *Journal of the American Statistical Association* **106**, 495, 1053–1066.

LÓPEZ-FIDALGO, J., TOMMASI, C. and TRANDAFIR, P. C. (2007). An optimal experimental design criterion for discriminating between non-normal models. *Journal of the Royal Statistical Society, Series B* **69**, 231–242.

MA, P., MAHONEY, M. W. and YU, B. (2015). A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research* **16**, 27, 861–911.

MACDOUGALL, J. (2006). *Analysis of Dose-Response Studies - $E_{\max}$ Model*. Springer, New York, TING, N. (ed.), p. 127–145.

PINHEIRO, J., BORNKAMP, B. and BRETZ, F. (2006). Design and analysis of dose-finding studies combining multiple comparisons and modeling procedures. *Journal of Biopharmaceutical Statistics* **16**, 639–656.

PÖTSCHER, B.M. (1991). Effects of model selection on inference. *Econometric Theory* **7**, 2, 163–185.

POWELL, M. J. D. (1994). *A direct search optimization method that models the objective and constraint functions by linear interpolation*. Dordrecht, HENNART, J.-P. and GOMEZ, S. (eds.), p. 51–67.

PUKELSHEIM, F. (2006). *Optimal Design of Experiments*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics. doi: 10.1137/1.9780898719109

QIAN, W., ROLLING, C. A., CHENG, G. and YANG, Y. (2019). On the forecast combination puzzle. *Econometrics* **7**(3).

RAFTERY, A. and ZHENG, Y. (2003). Discussion: Performance of bayesian model averaging. *Journal of the American Statistical Association* **98**, 931–938.

SCHORNING, K., BORNKAMP, B., BRETZ, F. and DETTE, H. (2016-09). Model selection versus model averaging in dose finding studies. *Statistics in Medicine* **35**, 22, 4021–4040.

SÉBASTIEN, B., HOFFMAN, D., RIGAUX, C., PELLISSIER, F. and MSIHID, J. (2016). Model averaging inconcentration-qt analyses. *Pharmaceutical Statistics* **15**, 6, 450–458.

SMITH, J. and WALLIS, K. F. (2009). A simple explanation of the forecast combination puzzle*. *Oxford Bulletin of Economics and Statistics* **71**, 3, 331–355.

STOCK, J. H. and WATSON, M. W. (2003). Forecasting output and inflation: The role of asset prices. *Journal of Economic Literature* **41**, 3, 788–829.

STOCK, J. H. and WATSON, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting* **23**, 6, 405–430.

TOMMASI, C. (2009). Optimal designs for both model discrimination and parameter estimation. *Journal of Statistical Planning and Inference* **139**, 4123–4132.

TOMMASI, C. and LÓPEZ-FIDALGO, J. (2010). Bayesian optimum designs for discriminating between models with any distribution. *Computational Statistics & Data Analysis* **54**, 1, 143–150.

UCINSKI, D. and BOGACKA, B. (2005). T-optimum designs for discrimination between two multiresponse dynamic models. *Journal of the Royal Statistical Society, Series B* **67**, 3–18.

VERRIER, D., SIVAPREGASSAM, S. and SOLENTE, A.-C. (2014). Dose-finding studies, mcp–mod, model selection, and model averaging: Two applications in the real world. *Clinical Trials* **11**, 4, 476–484. doi: doi: 10.1177/1740774514532723.

WANG, H., ZHANG, X. and ZOU, G. (2009). Frequentist model averaging estimation: a review. *Journal of Sysmtes Science and Complexity* **22**, 732.

WANG, H. (2019a). More efficient estimation for logistic regression with optimal subsamples. *Journal of Machine Learning Research* **20**, 132, 1–59.

WANG, H., YANG, M. and STUFKEN, J. (2019b). Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association* **114**, 525, 393–405.

WASSERMANN, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology* **44**, 92–107.

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1, 1–25.

YANG, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association* **96**, 454, 574–588.

YANG, Y. (2003). Regression with multiple candidate models: selecting or mixing?. *Statistica Sinica* **13**, 783–809.

YANG, Y. (2005). Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation. *Biometrika* **92**, 4, 937–950.

ZEN, M.-M. and TSAI, M.-H. (2002). Some criterion-robust optimal designs for the dual problem of model discrimination and parameter estimation. *Sankhya: The Indian Journal of Statistics* **64**, 322–338.

ZHANG, X., YU, D., ZOU, G. and LIANG, H. (2016). Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of the American Statistical Association* **111**, 516, 1775–1790.

Publisher's Note. Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix A: Technical Assumptions and Proofs

Assumptions

Following White (1982) we assume:

(A1) The random variables $Y_{ij}, i = 1, \ldots, k, j = 1, \ldots, n_i$ are independent. Furthermore, $Y_{i1}, \ldots, Y_{in_i}$ have a common distribution function with a measurable density $g(\cdot \mid x_i)$ with respect to a dominating measure ν.

(A2) The distribution function of each candidate model $s \in \{1, \ldots, r\}$ has a measurable density $f_s(\cdot \mid x, \theta_s)$ with respect to ν (for all $\theta_s \in \Theta_s$) that is continuous in θ_s.

(A3) For all $x \in \mathcal{X}$ the expectation $\mathbb{E}(\log(g(Y \mid x)))$ exists (where expectation is taken with respect to $g(\cdot \mid x)$) and for each candidate model the function $y \mapsto |\log f_s(y \mid x, \theta_s)|$ is dominated by a function that is integrable with respect to $g(\cdot \mid x)$ and does not depend on θ_s. Furthermore the Kullback-Leibler divergence (3.3) has a unique minimum $\theta^*_{s,g}(\xi)$ defined in Eq. 3.4 and $\theta^*_{s,g}(\xi)$ is an interior point of Θ_s.

(A4) For all $x \in \mathcal{X}$ the function $y \mapsto \frac{\partial \log f_s(y|x,\theta_s)}{\partial \theta_s}$ is a measurable function for all $\theta_s \in \Theta_s$ and continuously differentiable with respect to θ_s for all $y \in \mathbb{R}$.

(A5) The entries of the (matrix valued) functions $\frac{\partial^2 \log f_s(y|x,\theta_s)}{\partial \theta_s \partial \theta_s^\top}$, $\frac{\partial \log f_s(y|x,\theta_s)}{\partial \theta_s}$ $\big(\frac{\partial \log f_t(y|x,\theta_t)}{\partial \theta_t}\big)^\top$ are dominated by integrable functions with respect to $g(\cdot \mid x)$ for all $x \in \mathcal{X}$ and $\theta_s \in \Theta_s$.

(A6) The matrices $B_{ss}(\theta^*_s(\xi), \theta^*_s(\xi), \xi)$ and $A_s(\theta^*_s(\xi), \xi)$ in Eqs. 3.5 and 3.6 are nonsingular.

(A7) The functions $\theta_s \mapsto \mu_s(\theta_s)$ are once continuously differentiable.

Proof of Theorem 3.1..
By equation (A.2) in White (1982) we have

$$\sqrt{n}(\hat{\theta}_{n,s} - \theta^*_s(\xi)) + A_s^{-1}(\theta^*_s(\xi)) \frac{1}{\sqrt{n}} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{\partial \log f_s(Y_{ij} \mid x_i, \theta^*_s(\xi))}{\partial \theta_s} \xrightarrow{p} 0, \tag{A.1}$$

where $\xrightarrow{p}$ denotes convergence in probability (note that the matrix $A_s(\theta_s^*) = A_s(\theta_s^*, \xi)$ is nonsingular by assumption). An application of the multivariate central limit theorem now leads to

$$\frac{1}{\sqrt{n}} \begin{pmatrix} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{\partial \log f_1(Y_{ij}|x_i,\theta_1^*(\xi))}{\partial \theta_1} \\ \vdots \\ \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{\partial \log f_r(Y_{ij}|x_i,\theta_r^*(\xi))}{\partial \theta_r} \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \begin{pmatrix} B_{11} & \dots & B_{1r} \\ \vdots & \ddots & \vdots \\ B_{r1} & \dots & B_{rr} \end{pmatrix}\right), \tag{A.2}$$

where $B_{st} = B_{st}(\theta_s^*(\xi), \theta_t^*(\xi), \xi)$ is defined in Eq. 3.6. Combining (A.1) and (A.2) we obtain the weak convergence of the vector $\hat{\theta}_n = (\hat{\theta}_{n,1}^\top, \dots, \hat{\theta}_{n,r}^\top)^\top$, that is $\sqrt{n}(\hat{\theta}_n - \theta^*(\xi)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma)$, where $\Sigma = (\Sigma_{st})_{s,t=1,\dots,r}$ is a block matrix with entries $\Sigma_{st} = A_s^{-1}(\theta_s^*(\xi)) B_{st}(\theta_s^*(\xi), \theta_t^*(\xi)) A_t^{-1}(\theta_t^*(\xi))$ $(s,t = 1, \dots, r)$ and the vector $\theta_s^*(\xi)$ is given by $\theta_s^*(\xi) = (\theta_1^*(\xi)^\top, \dots, \theta_r^*(\xi)^\top)^\top$.

Next, we define for the parameter vector $\theta^\top = (\theta_1^\top, ..., \theta_r^\top) \in \mathbb{R}^{\sum_{s=1}^r p_s}$ the projection π_s by $\pi_s \theta := \theta_s$ and the vector $\tilde{\mu}(\theta) = \big(\mu_1(\pi_1\theta), \dots, \mu_r(\pi_r\theta)\big)^T$ with derivative

$$\mu'_\theta = \begin{pmatrix} \left(\frac{\partial \mu_1(\theta_1)}{\partial \theta_1}\right)^\top & 0 & \dots & & 0 \\ 0 & \left(\frac{\partial \mu_2(\theta_2)}{\partial \theta_2}\right)^\top & 0 & \dots & 0 \\ 0 & \dots & & 0 & \left(\frac{\partial \mu_r(\theta_r)}{\partial \theta_r}\right)^\top \end{pmatrix}. \tag{A.3}$$

An application of the Delta method shows that $\sqrt{n}(\tilde{\mu}(\hat{\theta}_n) - \tilde{\mu}(\theta^*(\xi))) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \mu'_{\theta^*(\xi)} \Sigma (\mu'_{\theta^*(\xi)})^\top\big)$. The assertion finally follows from the continuous mapping theorem observing the representation $\hat{\mu}_{\text{mav}} = (w_1, \dots, w_r)\, \tilde{\mu}(\hat{\theta}_n)$.

Proof of Theorem 3.2..

Throughout this proof we assume that integration and differentiation are interchangeable. Following the arguments in Pukelsheim (2006), Chapter 11, a Bayesian optimal design ξ^* for model averaging estimation of the parameter μ satisfies the inequality

$$-\int_{\mathcal{G}} D\Phi_{\text{mav}}^{\pi}(\xi, \mu_{\text{true}})(\xi_x - \xi^*) d\pi(g) \leq 0 \tag{A.4}$$

for all $x \in \mathcal{X}$, where $D\Phi_{\text{mav}}^{\pi}(\xi, \mu_{\text{true}})(\xi_x - \xi^*)$ denotes the directional derivative of the function Φ_{mav} evaluated in the optimal design ξ^* in direction $\xi_x - \xi^*$ and ξ_x denotes the Dirac measure at the point $x \in \mathcal{X}$.

To calculate the derivative we start with the derivative of the parameter $\theta^*_{s,g}(\xi)$ defined in Eq. 3.4 and define $\theta_{s,g}(\alpha) := \theta^*_{s,g}(\xi_\alpha)$ for $\xi_\alpha = \alpha\xi_x + (1-\alpha)\xi^*$. Note that $\theta_{s,g}(\alpha)$ is the solution of the equation

$$F_{s,g}(\alpha, \theta_s) = -\int\int g(y \mid t)\frac{\partial}{\partial\theta_s}\log f_s(y \mid t, \theta_s)dyd\xi_\alpha(t) = 0, \tag{A.5}$$

and that the derivatives of the left hand side are given by

$$\frac{\partial F_{s,g}}{\partial\alpha} = -\int\int g(y \mid t)\frac{\partial}{\partial\theta_s}\log f_s(y \mid t, \theta_s)dyd(\xi_x - \xi^*)(t), \tag{A.6}$$

$$\frac{\partial F_{s,g}}{\partial\theta_s} = -\int\int g(y \mid t)\frac{\partial^2}{\partial\theta_s\partial\theta_s^\top}\log f_s(y \mid t, \theta_s)dyd\xi_\alpha(t). \tag{A.7}$$

By the implicit function theorem we get $\frac{\partial\theta_{s,g}(\alpha)}{\partial\alpha} = -\big(\frac{\partial F_{s,g}}{\partial\theta_s}\big)^{-1}\frac{\partial F_{s,g}}{\partial\alpha}$ and hence

$$\begin{aligned}\frac{\partial}{\partial\alpha}\theta_{s,g}(\alpha)\Big|_{\alpha=0} &= -\Big(\int\int g(y \mid t)\frac{\partial^2}{\partial\theta_s\partial\theta_s^\top}\log f_s(y \mid t, \theta^*_{s,g}(\xi^*))dyd\xi^*(t)\Big)^{-1}\cdot \\ &\quad \int\int g(y \mid t)\frac{\partial}{\partial\theta_s}\log f_s(y \mid t, \theta^*_{s,g}(\xi^*))dyd(\xi_x - \xi^*)(t) = \theta'_{s,g}(\xi^*, x),\end{aligned}$$

where $\theta'_{s,g}(\xi^*, x)$ is defined in Eq. 3.16. Consider now the directional derivative of the matrix B_{st} defined in Eq. 3.6. An application of chain and product rule gives

$$\begin{aligned}&\frac{\partial B_{st}(\theta^*_{s,g}(\xi_\alpha), \theta^*_{t,g}(\xi_\alpha), \xi_\alpha)}{\partial\alpha}\Big|_{\alpha=0} \\ &= \Big(\int\int g(y \mid u)\frac{\partial}{\partial\alpha}\Big(\frac{\partial\log f_s(y \mid u, \theta^*_{s,g}(\xi_\alpha))}{\partial\theta_s}\Big(\frac{\partial\log f_t(y \mid u, \theta^*_{t,g}(\xi_\alpha))}{\partial\theta_t}\Big)^\top\Big)dyd\xi_\alpha(u) \\ &+ \int\int g(y \mid u)\frac{\partial\log f_s(y \mid u, \theta^*_{s,g}(\xi_\alpha))}{\partial\theta_s}\Big(\frac{\partial\log f_t(y \mid u, \theta^*_{t,g}(\xi_\alpha))}{\partial\theta_t}\Big)^\top dyd(\xi_x - \xi^*)(u)\Big)\Big|_{\alpha=0} \\ = \; & h'_{st,g}(\xi^*, x),\end{aligned}$$

where $h'_{st,g}(\xi^*, x)$ is defined in Eq. 3.18. In a similar way the derivative of the matrix A_s defined in Eq. 3.5 can be determined. First, using the chain rule, we observe that with $\theta'_{s,g}(\xi^*, x) = (\theta'_{s,g,1}(\xi^*, x), \cdots, \theta'_{s,g,p_s}(\xi^*, x))^\top$

$$\frac{\partial}{\partial\alpha}\left(\frac{\partial^2\log f_s(y \mid x, \theta^*_{s,g}(\xi_\alpha))}{\partial\theta_s\partial\theta_s^T}\right)\Bigg|_{\alpha=0} = D_s(\theta^*_{s,g}(\xi^*))(I_{p_s}\otimes\theta'_{s,g}(\xi^*, x)),$$

where D_s is defined in Theorem 3.2.

We now observe, that $\frac{\partial A_s(\theta^*_{s,g}(\xi_\alpha), \xi_\alpha)}{\partial\alpha}\big|_{\alpha=0} = h'_{s,g}(\xi^*, x)$, where $h'_{s,g}(\xi^*, x)$ is defined in Eq. 3.19.

Noting, that

$$\frac{\partial}{\partial\alpha}\frac{\partial\mu_s(\theta^*_{s,g}(\xi_\alpha))}{\partial\theta_s}\Big|_{\alpha=0} = \frac{\partial^2\mu_s(\theta^*_{s,g}(\xi^*))}{\partial\theta_s\partial\theta_s^\top}\theta'_{s,g}(\xi^*,x). \tag{A.8}$$

Equation 3.16 results by an application of the product rule and combination of the derivatives given above. Finally, we have

$$\begin{aligned}&\frac{\partial}{\partial\alpha}\Big(\sum_{s=1}^r w_s\mu_s(\theta^*_{s,g}(\xi)) - \mu_{\text{true}}\Big)^2\Big|_{\alpha=0}\\ &\quad= 2\Big(\sum_{s=1}^r w_s\mu_s(\theta^*_{s,g}(\xi^*)) - \mu_{\text{true}}\Big)\sum_{s=1}^r w_s\Big(\frac{\partial\mu_s(\theta^*_{s,g}(\xi^*))}{\partial\theta_s}\Big)^\top\theta'_{s,g}(\xi^*,x),\end{aligned}$$

and Eq. 3.15 follows.

The proof that there is equality in Eq. 3.15 for all support points of the optimal design ξ^* follows by a standard argument and the details are omitted for the sake of brevity.

Kira Alhorn
Kirsten Schorning
Fakultät Statistik, Technische
Universität Dortmund, Dortmund,
Germany

Holger Dette
Fakultät für Mathematik,
Ruhr-Universität Bochum, Bochum,
Germany
E-mail: holger.dette@ruhr-uni-bochum.de

Paper received: 24 January 2020; accepted 10 November 2020.