Optimal planning of factorial experiments in case of model contamination

Olga Egorova, Steven G. Gilmour

[Draft]

Contents

Abstract

1 Introduction

Designed experiments are commonly conducted in order to gain understanding of the impact that different parameters of interest make on the outcome; the quantitative measure of this impact allows comparisons as well as making interpretable conclusions regarding the scale and the pattern of the relationships between experimental factors and measured output.

While the exact true nature of that relationship remains unknown, some form of approximation is needed – and polynomial functions are able to provide any required precision for functions from a certain class of differentiability (Rudin, 1987). Response Surface Methodology (RSM, Box and Wilson (1951)) aims at optimising the approximated functions by fitting second-order polynomials, and higher desirable precision would require a polynomial of a higher order and, therefore, more experimental effort.

Whichever the chosen, 'primary', model is, planning a controlled intervention does rely on the approximating model assumptions in two main – and quite contradicting – directions. Firstly, it is desirable to ensure the goodness of what the model is essentially fitted for, that is, for example, the precision of model parameters and/or the prediction accuracy of the untested treatment combinations. The corresponding design aims are usually reflected in utilising well-known optimality criteria (D-, A-, G-,V-, etc.), which rely on the variance estimate, making the way it is obtained crucial. Secondly, treating the chosen model as the absolute truth, especially at the designing stage is at least too optimistic and, as in many cases fitting a more suitable, but complicated model might not be feasible due to various restrictions and limitations, could be dangerous for the results credibility. So having a particular model also means that at the stage of planning it is highly desirable to include some control over the model lack of fit and its effect on the inference.

We deal with such duality of model-dependence and accounting for its misspecification by developing compound optimality criteria, each constructed as a weighted combination of individual criterion functions, with the two main features:

1. Each corresponds to a specific desirable property: either accounting for an aim coming from trusting the model or mitigating the effects from its potential misfit. The relative importance of the components are reflected by the assigned weights – and we shall examine the performance of resulting optimal designs in terms of the individual criteria, and explore the role of weight allocation.

2. The concept of model-independent internal variance estimation, 'pure error' (Gilmour and Trinca (2012))) underlies each of the individual criteria – the most appropriate and sensible strategy in the case of possible model insufficiency.

The general spirit of this work aims at aligning with the concept of a good design, which, as summarised by Box and Draper (1987), should "make it possible to detect lack of fit" and "provide an internal estimate of error from replication", among other properties. Classical designs, like central composite ones (CCDs, Myers *et al.* (2009)), have been very popular in practice due to their geometrical properties.

We will focus on factorial experiments with a relatively small number of runs and the fitted model being a polynomial regression. Section 2 provides the background on the modelling, error estimation and fundamental individual criteria. Controlling the lack-of-fit and the bias arising from the model misspecification are introduced in Section 3, where they are combined with the primary model-driven ones in compound optimality criteria. Their adaptation to blocked experiments is described in Section 4, followed by the optimal design construction and variance analysis for experimental frameworks with restricted randomisation in Section 5. A series of examples are presented – examining dynamics across various optimal designs, shape of the constructed criteria and other properties, followed up by a Discussion in Section 6 with main conclusions and recommendations.

The R code used to search for the optimal designs are located at [github/R package link].

2 Model-dependent planning

Assuming a smooth enough relationship between k experimental factors $X_1, \ldots, X_k \in \Theta \subset \mathrm{R}^k$ and the response variable $Y = \eta(X_1, \ldots, X_k) \in \mathrm{R}$, a suitable polynomial model is chosen to fit data obtained from n experimental runs:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{1}$$

Here $\boldsymbol{X}$ is the $n \times p$ model matrix, $\boldsymbol{Y}$ is the $n \times 1$ vector of responses; $\boldsymbol{\beta}$ is the $p \times 1$ vector of parameters corresponding to the model terms and $\boldsymbol{\varepsilon}$ are independent normally distributed random error terms with constant variance: $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$.

Any inference based on building confidence regions and hypothesis testing following the model fitting relies on the error variance estimate $\hat{\sigma}^2$. The most common one is the mean square error: $\hat{\sigma}^2_{mse} = \text{Residual SS}/\nu$ (e.g. Draper and Smith (1998)), which is model-dependent with the residual degrees of freedom $\nu = n - p$ containing the number of model parameters.

The other one is 'pure' error, independent from the parametric model, which is derived from the further decomposition of the residual sum of squares into the 'pure' error and 'lack-of-fit' components: $\hat{\sigma}^2_{PE} = \text{Pure error SS}/(n-t)$, where t is the number of unique treatments (combinations of factors' levels) applied and $d = n - t$ is the pure error

degrees of freedom, that is the number of replicated points. In other words, the error is estimated as the mean square error from fitting the full treatment model:

$$\boldsymbol{Y} = \boldsymbol{X_t}\boldsymbol{\mu_t} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{X_t}$ is the $n \times t$ full treatment model matrix, in which the $(i,j)^{th}$ element is equal to 1 if treatment j is applied to the i^{th} unit, otherwise it is set to 0. Then the elements of the t-dimensional vector $\boldsymbol{\mu_t}$ are the mean effects of each treatment. The vector of errors $\boldsymbol{\varepsilon}$ comprises the between-unit variation, such that $\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$, $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$.

Many authors advocate the use of the 'pure' error estimate instead of the one pooled with the lack-of-fit part from the model (**??**): Cox (1958) recommends using it for the estimation unless there are no replicate treatments, while Draper and Smith (1998) argue for the reliability of 'pure' error and recommend aiming for the presence of replicates at the stage of planning. Atkinson *et al.* (2007) also state that "If lack-of-fit of the model is of potential interest, σ^2 is better estimated from replicate observations". Finally, the work by Gilmour and Trinca (2012) that inspired this research, comprises a thorough analysis in favour of estimating the error from the full treatment model which is true regardless of what function is used to approximate the relationship of interest.

However, model-dependency at the stage of experimental planning is reflected in searching for a design that optimises a criterion, that is a function of a design that translates a specific inference-driven objective. For example, among the most well-known, "alphabetic" optimality criteria, such as D-, C-, L-optimality and a series of others (Atkinson *et al.* (2007)) target the precision of parameters estimates in model (**??**); while others, like G- and V-optimality, deal with the prediction variance; traditionally these criteria are formulated based on the mean square error estimate. Gilmour and Trinca (2012) derived the alternative, model-independent pure-error based criteria, which would guarantee the presence of replicates in the resulting designs. Primary criteria were formulated for interval-based inferential properties: minimising the volume of a $(1-\alpha_{DP}) \times 100\%$ confidence region for the model parameters (DP-optimality) or the mean squared lengths of the $(1-\alpha_{LP}) \times 100\%$ confidence intervals for linear functions of the parameters' estimates' variances (LP-optimality):

$$\begin{aligned} DP: \ & (F_{p,d;1-\alpha_{DP}})^p |(\boldsymbol{X'X})^{-1}| \longrightarrow \ \min, \\ LP: \ & F_{1,d;1-\alpha_{LP}} \mathrm{tr}\{\boldsymbol{W}(\boldsymbol{X'X})^{-1}\} \longrightarrow \ \min, \end{aligned} \tag{3}$$

where $F_{df_1,df_2;1-\alpha}$ is the "upper α-point" of the F-distribution with df_1 and df_2 numerator and denominator degrees of freedom.

Combining multiple desirable objectives in the design can be fulfilled through constructing a compound criterion. This concept is based on the notion of design efficiency, which can be defined for any design matrix X and any criterion $C(X)$ as the ratio with respect to the best (e.g. minimum, without the loss of generality) value achieved at the

optimal design. For example, the DP-efficiency of design $\boldsymbol{X}$ is

$$\mathrm{Eff}_{DP}(X) = \frac{|\boldsymbol{X}_*'\boldsymbol{X}_*|^{1/p}/F_{p,d_*;1-\alpha_{DP}}}{|\boldsymbol{X}'\boldsymbol{X}|^{1/p}/F_{p,d;1-\alpha_{DP}}},$$

where $\boldsymbol{X}_*$ is the DP-optimum design with d_* pure error degrees of freedom. In this definition the power $1/p$ brings the efficiency to the scale of variances of model coefficients $\boldsymbol{\beta}_i, i = 1 \dots p$. The efficiency value may vary from 0 to 1 and is equal to 1 if and only if the design is optimal according to the criterion of interest.

The compound criterion to be maximised among all the possible designs is obtained then as weighted product of the individual criteria efficiencies $\mathrm{Eff}_1, \dots, \mathrm{Eff}_m$ with the corresponding weights $\kappa_1, \dots, \kappa_m$ (s.t. $\kappa_k > 0$ and $\sum_{k=1}^{m} \kappa_k = 1$):

$$\mathrm{Eff}_1^{\kappa_1}(\boldsymbol{X}) \times \mathrm{Eff}_2^{\kappa_2}(\boldsymbol{X}) \times \ldots \times \mathrm{Eff}_m^{\kappa_m}(\boldsymbol{X}) \to \max_{\boldsymbol{X}}. \tag{4}$$

The choice of weights is arbitrary in general, although is driven by prior knowledge of the experimenter, the objectives of a specific experiment and the components' interpretation. In the examples considered further in this work, we chose a set of weight allocations, most of which are classical design schemes for experiments with mixtures (Cornell, 2011).

Some of the alternative approaches of combining several objectives would include generating the Pareto optimal set of designs – the approach thoroughly described by Lu *et al.* (2011). Stallings and Morgan (2015)developed methodology for generalising eigenvalue-based criteria (e.g. A- and E-optimality) in a way that allows differing interest (expressed through the weights) among any set of estimable functions of the fitted model parameters. The introduced strategy reflects the aims of experimentation that are not traditionally accounted for but definitely are of interest.

3 Criteria Accounting for Model Uncertainty

Standard design optimality theory is developed under the assumption that the primary model (1) provides the most proper fit for the data: in many real applications this is quite a strong belief, and in reality we need to take into account at least the possibility that some misspecification is present at the planning stage.

In this work we consider the case when the fitted polynomial model with p parameters is nested within a larger model that is assumed to provide a better fit for the data:

$$\boldsymbol{Y} = \boldsymbol{X}_p\boldsymbol{\beta}_p + \boldsymbol{X}_q\boldsymbol{\beta}_q + \boldsymbol{\varepsilon}, \tag{5}$$

where $\boldsymbol{X}_q$ is an $n \times q$ extension of the primary model matrix containing extra q terms that we refer to as "potential terms" and that represent the fitted model disturbance,

with vector $\boldsymbol{\beta}_q$ denoting the corresponding parameters. They are not of any inferential interest, and, moreover, not all of them are necessarily estimable when the experiment is relatively small, i.e. $n < p+q$ – this is the case we mainly consider here. As usually, we do assume independent and normally distributed error terms: $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$; and even though the extended model is believed to potentially better fit for the data, it is not necessarily the one that should be used to obtain the estimates of σ^2 – the presence of model contamination is a strong arguement for using the model-independent replicate-based pure error estimate from the full treatment model (2).

3.1 Lack-of-fit criteria

Aiming towards controlling the magnitude and scope of the potential terms, we adapt Bayesian approach regarding the full model parameters, as was done by DuMouchel and Jones (1994). Diffuse prior shall be put on primary terms – with an arbitrary mean and a variance going to infinity, and the prior on potential terms is a normal distribution: $\boldsymbol{\beta}_q \sim \mathcal{N}(0, \boldsymbol{\Sigma}_0)$, where the prior variance scaled with respect to the error variation: $\boldsymbol{\Sigma}_0 = \sigma^2 \tau^2 \boldsymbol{I}_q$. Then, following the normality in model (5), the posterior distribution of the joint vector of coefficients $\boldsymbol{\beta} = [\boldsymbol{\beta}_p^T, \boldsymbol{\beta}_q^T]^T$ is multivariate normal (Koch (2007)):

$$\begin{gathered} \boldsymbol{\beta}|\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{b}, \boldsymbol{\Sigma}), \\ \text{where } \boldsymbol{b} = \boldsymbol{\Sigma}\boldsymbol{X}^T\boldsymbol{Y} \text{ and } \boldsymbol{\Sigma} = \sigma^2[\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{K}/\tau^2],^{-1} \boldsymbol{X} = [\boldsymbol{X}_p, \boldsymbol{X}_q], \\ \boldsymbol{K} = \begin{pmatrix} \mathbf{0}_{p\times p} & \mathbf{0}_{p\times q} \\ \mathbf{0}_{q\times p} & \boldsymbol{I}_{q\times q} \end{pmatrix}. \end{gathered}$$

The marginal posterior distribution of $\boldsymbol{\beta}_q$ is also multivariate normal with mean $\boldsymbol{b}_q$ – the last q elements of $\boldsymbol{b}$ and the covariance matrix $\boldsymbol{\Sigma}_{qq}$, that is the bottom right $q \times q$ submatrix of $\boldsymbol{\Sigma}$:

$$\begin{aligned} \boldsymbol{\Sigma}_{qq} = \sigma^2[(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{K}/\tau^2)^{-1}]_{[q,q]} &= \sigma^2 \begin{bmatrix} \boldsymbol{X}_p^T\boldsymbol{X}_p & \boldsymbol{X}_p^T\boldsymbol{X}_q \\ \boldsymbol{X}_q^T\boldsymbol{X}_p & \boldsymbol{X}_q^T\boldsymbol{X}_q + \boldsymbol{I}_q/\tau^2 \end{bmatrix}^{-1}_{[q,q]} \\ &= \sigma^2[\boldsymbol{X}_q^T\boldsymbol{X}_q + \boldsymbol{I}_q/\tau^2 - \boldsymbol{X}_q^T\boldsymbol{X}_p(\boldsymbol{X}_p^T\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^T\boldsymbol{X}_q]^{-1} \\ &= \sigma^2\left[\boldsymbol{L} + \boldsymbol{I}_q/\tau^2\right],^{-1} \end{aligned}$$

where $\boldsymbol{L} = \boldsymbol{X}_q^T\boldsymbol{X}_q - \boldsymbol{X}_q^T\boldsymbol{X}_p(\boldsymbol{X}_p^T\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^T\boldsymbol{X}_q$ is known in model-sensitivity design literature as the "dispersion matrix" (e.g. Goos *et al.* (2005)), which provides a point-wise measure of the distance between column vectors of $\boldsymbol{X}_q$ and linear subspace defined by the column vectors of the primary model matrix $\boldsymbol{X}_p$, that is how well each of the potential terms could be approximated by a linear span of the primary ones.

Reducing the primary model's lack-of-fit in the direction of the potential terms can be translated into a criterion function of the design by utilising the posterior distribution for $\boldsymbol{\beta}_q$ derived above and constructing a $(1-\alpha_{LoF}) \times 100\%$ posterior confidence region

which depends on the model matrices and the variance estimate s^2 on ν degrees of freedom (Draper and Smith, 1998):

$$(\boldsymbol{\beta}_2 - \boldsymbol{b}_2)'(\boldsymbol{L} + \boldsymbol{I}_q/\tau^2)(\boldsymbol{\beta}_2 - \boldsymbol{b}_2) \leq q s^2 F_{q,\nu;1-\alpha_{LoF}},$$

and minimising its volume is equivalent to

$$\left|\boldsymbol{L} + \boldsymbol{I}_q/\tau^2\right|^{-1/q} F_{q,d;1-\alpha_{LoF}} \longrightarrow \min, \tag{6}$$

which we refer to as "Lack-of-fit DP-criterion", that is directly related to (a) the lack-of-fit component in the Generalised D-optimality developed by Goos *et al.* (2005) – where the residual number of degrees of freedom ν does not depend on the design, and (b) DP-optimality (Gilmour and Trinca, 2012), with the F-quantile preserved from $\nu = d$ being the number of replicates in the design.

Another way of formulating a criterion to tackle the lack-of-fit that we consider in this work – minimising the average squared lengths of posterior confidence intervals for linear functions of $\boldsymbol{\beta}_q$ defined by matrix $\boldsymbol{J}$, – and we define the Lack-of-fit LP-criterion as mean of the squared lengths of the $(1 - \alpha_{LoF}) \times 100\%$ posterior confidence intervals for these linear functions:

$$\frac{1}{q}\text{trace}\left[\boldsymbol{J}\boldsymbol{J}^T\left(\boldsymbol{L} + \frac{\boldsymbol{I}_q}{\tau^2}\right)^{-1}\right] F_{1,d;1-\alpha_{LoF}} \longrightarrow \min. \tag{7}$$

This trace-based criterion is linked to the lack-of-fit part of the Generalised L-optimality (Goos *et al.*, 2005), and the pure error estimation approach retains the corresponding upper point of the F-distribution. Henceforth we mainly consider the case when $\boldsymbol{J}\boldsymbol{J}^T$ is diagonal, and the criterion above is reduced to weighted AP-optimality. In other words, the Lack-of-fit AP-criterion stands for the minimisation of the weighted average of the q-dimensional vector of the posterior confidence intervals' squared lengths for the potential parameters.

3.2 MSE-based criteria

Together with controlling the magnitude of model contamination, it is also desirable to "protect" the quality of inference that is to be drawn through fitting the primary model, from the potential presence of extra terms. From this point of view, the bias of the parameters' estimates $\hat{\boldsymbol{\beta}}_p$ would be of substantial interest; a natural way of evaluating their quality is the matrix of mean square error (Montepiedra and Fedorov, 1997), which is the $\mathcal{L}_2$-distance between the true and estimated values with respect to

the probability distribution measure of $\boldsymbol{Y}$ under the assumption of model (5):

$$\begin{aligned}\mathrm{MSE}(\hat{\boldsymbol{\beta}}_p|\boldsymbol{\beta}) =& \mathtt{E}_{\boldsymbol{Y}|\boldsymbol{\beta}}[(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_p)(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_p)^T] \\ =& \sigma^2(\boldsymbol{X}_p^T\boldsymbol{X}_p)^{-1} + \boldsymbol{A}\boldsymbol{\beta}_q\boldsymbol{\beta}_q^T\boldsymbol{A}^T, \end{aligned} \tag{8}$$

where $\boldsymbol{A} = (\boldsymbol{X}_p^T\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^T\boldsymbol{X}_q$ denotes the $p \times q$ alias matrix, which elements reflect the linearity scale of the relationship between the primary (rows) and potential (columns) terms.

We start constructing the determinant-based criterion that would correspond to the overall simultaneous minimisation of the bias above by taking log-determinant of the MSE matrix averaged across the prior distribution for $\boldsymbol{\beta}$:

$$\exp\{\mathtt{E}_{\boldsymbol{\beta}} \log(\det[\mathrm{MSE}(\hat{\boldsymbol{\beta}}_p|\boldsymbol{\beta})])\} \longrightarrow \min. \tag{9}$$

Using the matrix determinant lemma (Harville, 2006) and setting $\boldsymbol{M} = \boldsymbol{X}_p^T\boldsymbol{X}_p$ and $\tilde{\boldsymbol{\beta}}_2 = \boldsymbol{\beta}_2/\sigma$, the determinant and the applied logarithm in (9) can be decomposed:

$$\begin{aligned}&\det[\mathrm{MSE}(\hat{\boldsymbol{\beta}}_p|\boldsymbol{\beta})] \\ &= \det[\sigma^2\boldsymbol{M}^{-1} + \boldsymbol{M}^{-1}\boldsymbol{X}_p^T\boldsymbol{X}_q\boldsymbol{\beta}_q\boldsymbol{\beta}_q^T\boldsymbol{X}_q^T\boldsymbol{X}_p\boldsymbol{M}^{-1}] \\ &= \sigma^{2p}\det[\boldsymbol{M}^{-1} + \boldsymbol{M}^{-1}\boldsymbol{X}_p'\boldsymbol{X}_q\tilde{\boldsymbol{\beta}}_q\tilde{\boldsymbol{\beta}}_q^T\boldsymbol{X}_q^T\boldsymbol{X}_p\boldsymbol{M}^{-1}] \\ &= \sigma^{2p}\det[\boldsymbol{M}^{-1}]\det[1 + \tilde{\boldsymbol{\beta}}_q^T\boldsymbol{X}_q^T\boldsymbol{X}_p\boldsymbol{M}^{-1}\boldsymbol{M}\boldsymbol{M}^{-1}\boldsymbol{X}_p^T\boldsymbol{X}_q\tilde{\boldsymbol{\beta}}_q] \\ &= \sigma^{2p}\det[\boldsymbol{M}^{-1}](1 + \tilde{\boldsymbol{\beta}}_q^T\boldsymbol{X}_q^T\boldsymbol{X}_p\boldsymbol{M}^{-1}\boldsymbol{X}_p^T\boldsymbol{X}_q\tilde{\boldsymbol{\beta}}_q);\end{aligned}$$

$$\begin{aligned}&\log(\det[\mathrm{MSE}(\hat{\boldsymbol{\beta}}_p|\boldsymbol{\beta})]) \\ &= p\log\sigma^2 + \log(\det[\boldsymbol{M}^{-1}]) + \log(1 + \tilde{\boldsymbol{\beta}}_q^T\boldsymbol{X}_q^T\boldsymbol{X}_p\boldsymbol{M}^{-1}\boldsymbol{X}_p^T\boldsymbol{X}_q\tilde{\boldsymbol{\beta}}_q).\end{aligned}$$

The first summand does not depend on the design, so it will not be included in the criterion; the second one is the D-optimality function. Bringing the criterion in (9) to the scale of a single parameter, we obtain the following "MSE(D)-criterion":

$$|\boldsymbol{X}_p^T\boldsymbol{X}_p|^{-1/p}\exp\{\mathtt{E}_{\tilde{\boldsymbol{\beta}}_q}\log(1 + \tilde{\boldsymbol{\beta}}_q^T\boldsymbol{X}_q^T\boldsymbol{X}_p(\boldsymbol{X}_p^T\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^T\boldsymbol{X}_q\tilde{\boldsymbol{\beta}}_q)\}^{1/p} \longrightarrow \min. \tag{10}$$

Due to the obvious lack of information regarding $\tilde{\boldsymbol{\beta}}_q$, the expectation in the second term needs to be evaluated numerically. Expressing the prior variance of $\boldsymbol{\beta}_q$ as a scaled error variance $\boldsymbol{\beta}_q \sim \mathcal{N}(\boldsymbol{0}, \tau^2\sigma^2\boldsymbol{I}_q)$ means that $\tilde{\boldsymbol{\beta}}_q \sim \mathcal{N}(\boldsymbol{0}, \tau^2\boldsymbol{I}_q)$, and that, quite conveniently, its prior distribution does not depend on the unknown σ^2. Then a primitive Monte-Carlo can be used to evaluate that term: drawing a sample of large size N from the prior, and approximating the expectation above by the average across the sampled

values of $\tilde{\boldsymbol{\beta}}_{q_i}$, $i = 1, \ldots, N$:

$$\mathtt{E}_{\tilde{\boldsymbol{\beta}}_q} \log(1 + \tilde{\boldsymbol{\beta}}_q^T \boldsymbol{X}_q^T \boldsymbol{X}_p \boldsymbol{M}^{-1} \boldsymbol{X}_p^T \boldsymbol{X}_q \tilde{\boldsymbol{\beta}}_q) \approx \frac{1}{N} \sum_{i=1}^{N} \log(1 + \tilde{\boldsymbol{\beta}}_{q_i}^T \boldsymbol{X}_q^T \boldsymbol{X}_p \boldsymbol{M}^{-1} \boldsymbol{X}_p^T \boldsymbol{X}_q \tilde{\boldsymbol{\beta}}_{q_i}).$$

One of the alternatives to this computationally demanding approach is to use the point prior for $\boldsymbol{\beta}_q$, that is setting $\boldsymbol{\beta}_q = \sigma\tau\mathbf{1}_q$ (where $\mathbf{1}_q$ is a q-dimensional vector of 1s), which is the standard deviation of the initial normal prior. Then $\tilde{\boldsymbol{\beta}}_q = \tau\mathbf{1}_q$ with probability 1 and the expectation above becomes:

$$\begin{aligned}\mathtt{E}_{\tilde{\boldsymbol{\beta}}_q} \log(1 + \tilde{\boldsymbol{\beta}}_q^T \boldsymbol{X}_q^T \boldsymbol{X}_p \boldsymbol{M}^{-1} \boldsymbol{X}_p^T \boldsymbol{X}_q \tilde{\boldsymbol{\beta}}_q) &\approx \log(1 + \tau^2 \mathbf{1}_q^T \boldsymbol{X}_q^T \boldsymbol{X}_p \boldsymbol{M}^{-1} \boldsymbol{X}_p^T \boldsymbol{X}_q \mathbf{1}_q) \\ &= \log(1 + \tau^2 \sum_{i,j=1}^{q} [\boldsymbol{X}_q^T \boldsymbol{X}_p \boldsymbol{M}^{-1} \boldsymbol{X}_p^T \boldsymbol{X}_q][i,j]),\end{aligned}$$

with summation of matrix elements taking considerably less computational time even for large q.

To infer the trace-based form of the criterion that would minimise the mean squared bias, we use the pseudo-Bayesian approach of setting priors on the model's parameters to calculating the trace function of the MSE matrix (8):

$$\begin{aligned}\mathtt{E}_{\boldsymbol{\beta}_q} \mathrm{trace}[\mathrm{MSE}(\hat{\boldsymbol{\beta}}_p \boldsymbol{\beta}_q)] &= \mathrm{trace}[\mathtt{E}_{\boldsymbol{\beta}_q} \mathrm{MSE}(\hat{\boldsymbol{\beta}}_p | \boldsymbol{\beta}_q)] \\ &= \mathrm{trace}[\sigma^2 (\boldsymbol{X}_p^T \boldsymbol{X}_p)^{-1} + \mathtt{E}_{\boldsymbol{\beta}_q} (\boldsymbol{A}\boldsymbol{\beta}_q \boldsymbol{\beta}_q^T \boldsymbol{A}^T)] \\ &= \mathrm{trace}[\sigma^2 (\boldsymbol{X}_p^T \boldsymbol{X}_p)^{-1} + \sigma^2\tau^2 \boldsymbol{A}\boldsymbol{A}^T] \\ &= \sigma^2 \mathrm{trace}[(\boldsymbol{X}_p^T \boldsymbol{X}_p)^{-1} + \tau^2 \boldsymbol{A}\boldsymbol{A}^T] \\ &= \sigma^2 [\mathrm{trace}\{(\boldsymbol{X}_p^T \boldsymbol{X}_p)^{-1}\} + \tau^2 \mathrm{trace} \boldsymbol{A}\boldsymbol{A}^T].\end{aligned}$$

The operations of calculating trace and expectation are commutative, hence there is no necessity of any additional numerical evaluations, and in this case of the trace-based criterion using the point prior for $\boldsymbol{\beta}_q$ defined earlier would lead to the same resulting function. By minimising the whole function above, we simultaneously minimise both the sum of primary terms' variance and the expected squared norm of the bias vector in the direction of the potential terms. The second part here contains the scaling parameter τ^2 which regulates the magnitude of the potential terms' variation relatively to the error variance.

We formally define the "MSE(L)-criterion" as

$$\frac{1}{p} \mathrm{trace}\{(\boldsymbol{X}_p^T \boldsymbol{X}_p)^{-1} + \tau^2 \boldsymbol{A}\boldsymbol{A}^T\} \longrightarrow \text{ min.} \tag{11}$$

Compound criteria

The compound criterion in a general form is constructed to account for the three main objectives: precision of the primary model parameters, control for the lack-of-fit and minimising the inferential bias from the potential model contamination – as a weighted product of efficiencies (4) with respect to DP-/LP-criterion, Lack-of-Fit functions(6, 7), and the MSE-based ones in (10, 11).

$$\begin{aligned}
&\left[\left|\boldsymbol{X}_p^T\boldsymbol{X}_p\right|^{-1/p} F_{p,d;1-\alpha_{DP}}\right]^{\kappa_{DP}} \times \\
&\left[\left|\boldsymbol{L}+\frac{\boldsymbol{I}_q}{\tau^2}\right|^{-1/q} F_{q,d;1-\alpha_{LoF}}\right]^{\kappa_{LoF}} \times \\
&\left[|\boldsymbol{X}_p^T\boldsymbol{X}_p|^{-1}\exp\left(\mathtt{E}_{\tilde{\boldsymbol{\beta}}_q}\log(1+\tilde{\boldsymbol{\beta}}_q^T\boldsymbol{X}_q^T\boldsymbol{X}_p\boldsymbol{M}^{-1}\boldsymbol{X}_p^T\boldsymbol{X}_q\tilde{\boldsymbol{\beta}}_q)\right)\right]^{\kappa_{MSE}/p} \longrightarrow \min. \qquad (12)
\end{aligned}$$

This compound criterion is later referred to as compound "MSE-DP-criterion", where α_{DP} and α_{LoF} denote the confidence intervals' probability levels for the primary $\boldsymbol{\beta}_p$ and potential coefficients $\boldsymbol{\beta}_q$ – in the DP- and LoF-DP elementary criteria accordingly. As mentioned before, non-negative weights κ_X that sum up to 1 define the compound criterion and are chosen to reflect the experimenter's priorities.

Similarly, we obtain the compound "MSE-LP-criterion" by joining the LP criterion with trace-based Lack-of-fit (7)) and MSE components :

$$\begin{aligned}
&\left[\frac{1}{p}\text{trace}(\boldsymbol{W}\boldsymbol{X}_p^T\boldsymbol{X}_p)^{-1}F_{1,d;1-\alpha_{LP}}\right]^{\kappa_{LP}} \times \\
&\left[\frac{1}{q}\text{trace}\left(\boldsymbol{L}+\frac{\boldsymbol{I}_q}{\tau^2}\right)^{-1}F_{1,d;1-\alpha_{LoF}}\right]^{\kappa_{LoF}} \times \\
&\left[\frac{1}{p}\text{trace}\{(\boldsymbol{X}_p^T\boldsymbol{X}_p)^{-1}+\tau^2\boldsymbol{A}\boldsymbol{A}^T\}\right]^{\kappa_{MSE}} \longrightarrow \min. \qquad (13)
\end{aligned}$$

Confidence levels α play similar role here, but they do not have to be the same as in the determinant-based criterion. Moreover, it would be sensible to take into account the multiple testing corrections, as we are dealing with minimising multiple confidence intervals rather than with a single region's volume. The next subsection explores an illustrative example where optimal designs were searched using the compound optimality criteria constructed above.

3.3 Example

Here we will study the optimal designs in terms of the criteria (12) and (13) in the framework a factorial experiment, with 3 factors, each is at three levels. The small number of runs ($n = 16$) allows estimation of the full-second order polynomial model

($p = 10$), but we assume that the extended model, potentially providing a better fit, contains also all third-order terms (linear-by-linear-by-linear and quadratic-by-linear interactions), $q = 30$ of them in total.

The intercept is a nuisance parameter, and so the criteria are adapted in such a way that the full information matrix in the DP- and LP-components and in the first part of the MSE-based components is replaced by the one excluding the intercept $\boldsymbol{M}_0 = \boldsymbol{X}_{p-1}^T \boldsymbol{Q}_0 \boldsymbol{X}_{p-1}$, where $\boldsymbol{X}_{p-1}$ is the model matrix without the intercept, $\boldsymbol{Q}_0 = \boldsymbol{I}_n - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}'$. Otherwise the procedure of obtaining the $MSE(D)$- and $MSE(L)$-based components remains the same, and the MSE-based criteria functions that have been amended according to the intercept exclusion are referred to as MSE-DPs and MSE-LPs (similar to DPs and LPs from Gilmour and Trinca (2012)):

$$
\begin{aligned}
\text{MSE-DPs: } & \left[\left| \boldsymbol{X}_{p-1}^T \boldsymbol{Q}_0 \boldsymbol{X}_{p-1} \right|^{-1/(p-1)} F_{p-1,d;1-\alpha_{DP}} \right]^{\kappa_{DP}} \times \\
& \left[\left| \boldsymbol{L} + \frac{\boldsymbol{I}_q}{\tau^2} \right|^{-1/q} F_{q,d;1-\alpha_{LoF}} \right]^{\kappa_{LoF}} \times \\
& \left[|\boldsymbol{X}_{p-1}' \boldsymbol{Q}_0 \boldsymbol{X}_{p-1}|^{-1} \exp\left(\mathrm{E}_{\tilde{\boldsymbol{\beta}}_q} \log(1 + \tilde{\boldsymbol{\beta}}_q^T \boldsymbol{X}_q^T \boldsymbol{Q}_0 \boldsymbol{X}_{p-1} \boldsymbol{M}_0^{-1} \boldsymbol{X}_{p-1}^T \boldsymbol{Q}_0 \boldsymbol{X}_q \tilde{\boldsymbol{\beta}}_q) \right) \right]^{\kappa_{MSE}/(p-1)}, \\
\text{MSE-LPs: } & \left[\frac{1}{p-1} \mathrm{trace}(\boldsymbol{W} \boldsymbol{X}_{p-1}^T \boldsymbol{Q}_0 \boldsymbol{X}_{p-1})^{-1} F_{1,d;1-\alpha_{LP}} \right]^{\kappa_{LP}} \times \\
& \left[\frac{1}{q} \mathrm{trace} \left(\boldsymbol{L} + \frac{\boldsymbol{I}_q}{\tau^2} \right)^{-1} F_{1,d;1-\alpha_{LoF}} \right]^{\kappa_{LoF}} \times \\
& \left[\frac{1}{p-1} \mathrm{trace}[\boldsymbol{M}^{-1} + \tau^2 \boldsymbol{A}\boldsymbol{A}^T]_{[p-1,p-1]} \right]^{\kappa_{MSE}}.
\end{aligned}
$$

In the third component of the MSE-LP-criterion $[\boldsymbol{M}^{-1} + \tau^2 \boldsymbol{A}\boldsymbol{A}^T]_{[p-1,p-1]}$ stands for the submatrix corresponding to the parameters of interest, i.e. the first column and row removed.

Considering two values of the variance scaling parameter $\tau^2 = 1$ and $\tau^2 = 1/q$, for each compound criterion we obtain two sets of optimal designs, their summaries are given in Tables 1 and 2. Every row corresponds to a design that has been obtained as optimal according to the compound criterion defined by the combination of weights κ-s. We explore the distribution of degrees of freedom between the pure error and lack-of-fit components in the designs and at the optimal designs' efficiencies with respect to the individual criteria that are given in the last columns. Optimal designs were obtained using the point exchange algorithm (Fedorov (1972)), with 500 random starts; the MSE(D)-part of the compound criterion was estimated using the MC sampling, and this is the most time consuming part of the computations. When this presents a considerable challenge, we can recommend the previously mentioned alternative of imputing the point prior values of $\tilde{\boldsymbol{\beta}}_{\boldsymbol{q}}$. The resulting losses in the efficiencies are quite small, and time savings are substantial – the illustration on this example is presented in Appendix A.1.

[the Results of the smaller example are in preparation]

Table 1: Properties of MSE-DP-optimal designs

	Criteria, $\tau^2 = 1$			**DoF**		**Efficiency,%**					
	DP	**LoF(DP)**	**MSE(D)**	**PE**	**LoF**	**DP**	**LoF(DP)**	**MSE(D)**	**LP**	**LoF(LP)**	**MSE(L)**
1	1	0	0	18	1	100.00	47.77	91.05	96.38	94.92	10.70
2	0	1	0	8	11	43.70	100.00	54.74	0.75	89.99	2.08
3	0	0	1	0	19	0.00	0.00	100.00	0.00	0.00	22.32
4	0.5	0.5	0	11	8	78.50	87.61	88.56	73.88	98.85	17.66
5	0.5	0	0.5	15	4	97.26	56.51	93.77	97.55	50.04	12.74
6	0	0.5	0.5	8	11	64.72	96.84	87.53	57.04	36.17	29.33
7	1/3	1/3	1/3	10	9	79.45	84.14	93.23	81.06	43.42	16.71
8	0.5	0.25	0.25	13	6	93.38	64.35	95.55	95.76	48.58	14.77
9	0.25	0.5	0.25	9	10	69.52	95.76	87.36	63.13	40.41	25.46
	Criteria, $\tau^2 = 1/q$			**DoF**		**Efficiency,%**					
	DP	**LoF(DP)**	**MSE(D)**	**PE**	**LoF**	**DP**	**LoF(DP)**	**MSE(D)**	**LP**	**LoF(LP)**	**MSE(L)**
1	1	0	0	18	1	100.00	94.41	90.60	96.42	98.36	44.94
2	0	1	0	16	3	39.66	100.00	37.95	0.13	100.00	0.12
3	0	0	1	0	19	0.00	0.00	100.00	0.00	0.00	77.97
4	0.5	0.5	0	18	1	100.00	94.41	90.60	96.42	98.36	44.94
5	0.5	0	0.5	17	2	97.31	91.28	93.98	96.21	94.32	50.53
6	0	0.5	0.5	15	4	96.10	92.31	93.29	99.48	95.09	57.05
7	1/3	1/3	1/3	18	1	100.00	94.41	90.60	96.42	98.36	44.94
8	0.5	0.25	0.25	18	1	100.00	94.41	90.60	96.42	98.36	44.94
9	0.25	0.5	0.25	18	1	99.96	94.33	90.65	96.24	98.31	44.84

Table 2: Properties of MSE(L)-optimal designs

	Criteria, $\tau^2 = 1$			**DoF**		**Efficiency,%**					
	LP	**LoF(LP)**	**MSE(L)**	**PE**	**LoF**	**DP**	**LoF(DP)**	**MSE(D)**	**LP**	**LoF(LP)**	**MSE(L)**
1	1	0	0	16	3	97.54	53.86	92.54	100.00	96.87	12.08
2	0	1	0	13	6	35.43	81.99	36.72	0.00	100.00	0.00
3	0	0	1	4	15	18.67	38.43	51.84	11.73	34.79	100.00
4	0.5	0.5	0	15	4	95.14	60.37	92.78	99.80	98.12	13.99
5	0.5	0	0.5	12	7	77.77	72.05	84.91	81.10	98.72	25.19
6	0	0.5	0.5	9	10	36.80	70.91	51.12	28.13	91.60	83.52
7	1/3	1/3	1/3	11	8	69.53	73.16	79.71	70.59	97.47	27.98
8	0.5	0.25	0.25	12	7	77.20	72.83	84.44	81.47	98.80	23.88
9	0.25	0.5	0.25	12	7	70.90	69.80	78.15	72.16	98.49	26.19
	Criteria, $\tau^2 = 1/q$			**DoF**		**Efficiency,%**					
	LP	**LoF(LP)**	**MSE(L)**	**PE**	**LoF**	**DP**	**LoF(DP)**	**MSE(D)**	**LP**	**LoF(LP)**	**MSE(L)**
1	1	0	0	16	3	97.54	92.13	92.18	100.00	95.61	52.55
2	0	1	0	16	3	39.66	100.00	37.95	0.13	100.00	0.12
3	0	0	1	3	16	0.77	0.93	83.48	0.02	0.01	100.00
4	0.5	0.5	0	17	2	96.87	94.29	89.84	97.97	97.80	51.17
5	0.5	0	0.5	12	7	79.66	87.18	86.32	84.81	88.22	79.60
6	0	0.5	0.5	13	6	76.46	89.87	80.78	79.23	91.48	79.11
7	1/3	1/3	1/3	13	6	81.53	90.09	85.52	85.96	91.56	76.65
8	0.5	0.25	0.25	15	4	90.60	91.94	88.43	95.09	94.75	63.82
9	0.25	0.5	0.25	15	4	84.12	92.88	83.39	87.86	95.51	72.66

The observed relationships between the components of the criteria provide some ideas on how the designs obtained as optimal with respect to the various allocations of weights differ, what they have in common, and how the components interact when the weights are reallocated. A compromise can be found between the criterion components corresponding to the properties of the primary terms and the components responsible for the reducing the negative impact from the assumed potential model misspecification. Thereby, the careful choice of the prior parameters (τ^2 in this case) seems to be of a certain importance as well.

4 Blocked Experiments

In some experiments, where the number of runs is relatively large, and/or the variability between units is quite high, experimental units are allocated in blocks such that within each block the units are similar in some way, e.g. the runs carried out within the same day. Such formalisation contributes to controlling the variability by separating variation coming from the difference between blocks and the variability arising from within the blocks (Bailey, 2008).

Under the assumption of additivity of b fixed block effects the polynomial model can be presented as:

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta}_B + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{14}$$

which, in addition to the same terms as in (1) comprises $\boldsymbol{Z}$ – the $n \times b$ matrix, such that its $(i,j)^{th}$ element is equal to 1 if unit i is in block j and to 0 otherwise, and $\boldsymbol{\beta}_B$ is the vector of block effects.

The full information matrix has the form $\boldsymbol{M_B} = \begin{pmatrix} \boldsymbol{Z'Z} & \boldsymbol{Z'X} \\ \boldsymbol{X'Z} & \boldsymbol{X'X} \end{pmatrix}$.

Using the rules of inverting blocked matrices (Harville, 2006), we can isolate the variance of the polynomial coefficients' estimates:

$$\begin{aligned} &\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{M_B}^{-1})_{22} = \sigma^2(\boldsymbol{X'QX}),^{-1} \\ &\text{where } \boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z'Z})^{-1}\boldsymbol{Z'}. \end{aligned}$$

The DP- and LP-criteria become DP_S and LP_S in the context of a blocked experiment, and they can be straightforwardly defined as

$$DP_S : (F_{p,d_B;1-\alpha_{DP}})^p |(\boldsymbol{X'QX})^{-1}| \rightarrow \min, \tag{15}$$

$$LP_S : F_{1,d_B;1-\alpha_{LP}} \mathrm{trace}\{\boldsymbol{W}(\boldsymbol{X'QX})^{-1}\} \rightarrow \min.. \tag{16}$$

The number of pure error degrees of freedom is now calculated as $d_B = n - \mathrm{rank}[\boldsymbol{Z} : \boldsymbol{T}]$, where $\boldsymbol{T}$ is the $n \times t$ matrix whose elements indicate the treatments (Gilmour and Trinca, 2012), providing the number of replications after subtracting the ones taken for the estimation of block contrasts.

To adapt the derivation of the lack-of-fit and MSE-based criteria to the blocked experiments, we start by formulating the model comprising both primary terms and possible contamination in the form of potential terms, now for blocked experiments:

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta}_B + \boldsymbol{X}_p\boldsymbol{\beta}_p + \boldsymbol{X}_q\boldsymbol{\beta}_q + \boldsymbol{\varepsilon}.$$

Denote the $n \times (b+p)$ model matrix of the block and primary terms by $\tilde{\boldsymbol{X}}_p = [\boldsymbol{Z}, \boldsymbol{X}_p]$ and let $\tilde{\boldsymbol{\beta}}_p = [\boldsymbol{\beta}_B, \boldsymbol{\beta}_p]'$ be the joint vector of fixed block effects and primary model terms,

and by $\hat{\tilde{\boldsymbol{\beta}}}_p$ we denote the vector of the corresponding estimates. It is worth noting that the number of primary terms p does not include the intercept, as it is aliased with the block effects.

4.1 Lack-of-fit criteria

The information matrix for the model (14), up to a multiple of σ^2, is:

$$\boldsymbol{M_B} = \begin{pmatrix} \boldsymbol{Z}^T\boldsymbol{Z} & \boldsymbol{Z}^T\boldsymbol{X}_p & \boldsymbol{Z}^T\boldsymbol{X}_q \\ \boldsymbol{X}_p^T\boldsymbol{Z} & \boldsymbol{X}_p^T\boldsymbol{X}_p & \boldsymbol{X}_p^T\boldsymbol{X}_q \\ \boldsymbol{X}_q^T\boldsymbol{Z} & \boldsymbol{X}_q^T\boldsymbol{X}_p & \boldsymbol{X}_q^T\boldsymbol{X}_q + \boldsymbol{I}_q/\tau^2 \end{pmatrix} = \begin{pmatrix} \tilde{\boldsymbol{X}}_p^T\tilde{\boldsymbol{X}}_p & \tilde{\boldsymbol{X}}_p^T\boldsymbol{X}_q \\ \boldsymbol{X}_q^T\tilde{\boldsymbol{X}}_p & \boldsymbol{X}_q^T\boldsymbol{X}_q + \boldsymbol{I}_q/\tau^2 \end{pmatrix}.$$

Assuming the same normal prior on $\boldsymbol{\beta}_q \sim \mathcal{N}(\boldsymbol{0}, \tau^2\sigma^2\boldsymbol{I}_q)$, we can construct the variance-covariance matrix corresponding to the potential terms, which would be the inverse of the lower right submatrix of $\boldsymbol{M_B}$: $\tilde{\boldsymbol{\Sigma}}_{qq} = \sigma^2[\boldsymbol{M}_B^{-1}]_{22}$.

$$\begin{aligned} \tilde{\boldsymbol{\Sigma}}_{qq} &= \sigma^2([\boldsymbol{M_B}]_{22} - [\boldsymbol{M_B}]_{21}([\boldsymbol{M_B}]_{11})^{-1}[\boldsymbol{M_B}]_{12})^{-1} \\ &= \sigma^2(\boldsymbol{X}_q^T\boldsymbol{X}_q + \boldsymbol{I}_q/\tau^2 - \boldsymbol{X}_q^T\tilde{\boldsymbol{X}}_p(\tilde{\boldsymbol{X}}_p^T\tilde{\boldsymbol{X}}_p)^{-1}\tilde{\boldsymbol{X}}_p^T\boldsymbol{X}_q)^{-1} \\ &= \sigma^2\left(\tilde{\boldsymbol{L}} + \frac{\boldsymbol{I}_q}{\tau^2}\right), \text{ where } \tilde{\boldsymbol{L}} = \boldsymbol{X}_q^T\boldsymbol{X}_q - \boldsymbol{X}_q^T\tilde{\boldsymbol{X}}_p(\tilde{\boldsymbol{X}}_p^T\tilde{\boldsymbol{X}}_p)^{-1}\tilde{\boldsymbol{X}}_p^T\boldsymbol{X}_q. \end{aligned}$$

Therefore, the lack-of-fit criteria in (6) and (7) are adjusted for blocked experiments by replacing the primary terms matrix $\boldsymbol{X}_p$ by the extended matrix $\tilde{\boldsymbol{X}}_p$ and the dispersion matrix $\boldsymbol{L}$ – by $\tilde{\boldsymbol{L}}$ as obtained above.

4.2 MSE-based criteria

As for the MSE-based measure of the shift in the primary terms estimates, we first consider the overall mean square matrix

$$\begin{aligned} \text{MSE}(\hat{\tilde{\boldsymbol{\beta}}}_p|\tilde{\boldsymbol{\beta}}) =& \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{\beta}}[(\hat{\tilde{\boldsymbol{\beta}}}_p - \tilde{\boldsymbol{\beta}}_p)(\hat{\tilde{\boldsymbol{\beta}}}_p - \tilde{\boldsymbol{\beta}}_p)'] \\ =& \sigma^2(\tilde{\boldsymbol{X}}_p'\tilde{\boldsymbol{X}}_p)^{-1} + \tilde{\boldsymbol{A}}\boldsymbol{\beta}_q\boldsymbol{\beta}_q^T\tilde{\boldsymbol{A}}^T, \end{aligned} \tag{17}$$

with $\tilde{\boldsymbol{A}} = (\tilde{\boldsymbol{X}}_p^T \tilde{\boldsymbol{X}}_p)^{-1} \tilde{\boldsymbol{X}}_p'^T \boldsymbol{X}_q$ being the alias matrix, and its partition with respect to block and primary effects:

$$
\begin{aligned}
&\mathrm{E}_{\boldsymbol{Y}|\boldsymbol{\beta}}[(\hat{\tilde{\boldsymbol{\beta}}}_p - \tilde{\boldsymbol{\beta}}_p)(\hat{\tilde{\boldsymbol{\beta}}}_p - \tilde{\boldsymbol{\beta}}_p)^T] = \\
&\mathrm{E}_{\boldsymbol{Y}|\boldsymbol{\beta}}\{[\hat{\tilde{\beta}}_{p1} - \tilde{\beta}_{p1}, \ldots, \hat{\tilde{\beta}}_{pb} - \tilde{\beta}_{pb}, \hat{\tilde{\beta}}_{pb+1} - \tilde{\beta}_{pb+1}, \ldots, \hat{\tilde{\beta}}_{pb+p} - \tilde{\beta}_{pb+p}] \times \\
&[\hat{\tilde{\beta}}_{p1} - \tilde{\beta}_{p1}, \ldots, \hat{\tilde{\beta}}_{pb} - \tilde{\beta}_{pb}, \hat{\tilde{\beta}}_{pb+1} - \tilde{\beta}_{pb+1}, \ldots, \hat{\tilde{\beta}}_{pb+p} - \tilde{\beta}_{pb+p}]^T\} = \\
&\mathrm{E}_{\boldsymbol{Y}|\boldsymbol{\beta}}\{[\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_B, \hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_p][\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_B, \hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_p]^T\} = \\
&\begin{bmatrix} \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_B)(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_B)^T & \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_B)(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_p)^T \\ \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_p)(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_B)^T & \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_p)(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_p)^T \end{bmatrix}.
\end{aligned}
$$

The part corresponding to the bias of the primary terms $\boldsymbol{\beta}_p$ is the lower right $p \times p$ submatrix, and we can extract it from the MSE expression in (17). The respective submatrix of the first summand is

$$
[\sigma^2(\tilde{\boldsymbol{X}}_p^T \tilde{\boldsymbol{X}}_p)^{-1}]_{22} = \sigma^2(\boldsymbol{X}_p^T \boldsymbol{Q} \boldsymbol{X}_p)^{-1}, \text{ where } \boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T
$$

Using the matrix inversion rule for block matrices (Harville, 2006), we now consider $\tilde{\boldsymbol{A}}$:

$$
\begin{aligned}
\tilde{\boldsymbol{A}} = &\left(\begin{bmatrix} \boldsymbol{Z}^T \\ \boldsymbol{X}_p^T \end{bmatrix} \begin{bmatrix} \boldsymbol{Z} & \boldsymbol{X}_p \end{bmatrix} \right)^{-1} \begin{bmatrix} \boldsymbol{Z}^T \\ \boldsymbol{X}_p^T \end{bmatrix} \boldsymbol{X}_q = \begin{pmatrix} \boldsymbol{Z}^T\boldsymbol{Z} & \boldsymbol{Z}^T\boldsymbol{X}_p \\ \boldsymbol{X}_p^T\boldsymbol{Z} & \boldsymbol{X}_p^T\boldsymbol{X}_p \end{pmatrix}^{-1} \begin{bmatrix} \boldsymbol{Z}^T \\ \boldsymbol{X}_p^T \end{bmatrix} \boldsymbol{X}_q = \\
&\begin{bmatrix} (\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{Z})^{-1} & -(\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{X}_p(\boldsymbol{X}_p^T\boldsymbol{X}_p)^{-1} \\ -(\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_p^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1} & (\boldsymbol{X}_1^T\boldsymbol{Q}\boldsymbol{X}_p)^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{Z}^T \\ \boldsymbol{X}_p^T \end{bmatrix} \boldsymbol{X}_q = \\
&\begin{bmatrix} (\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{X}_q \\ (\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_q \end{bmatrix},
\end{aligned}
$$

where $\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{X}_p(\boldsymbol{X}_p^T\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^T$; $\boldsymbol{Z}\boldsymbol{Z}'$, $\boldsymbol{X}_1'\boldsymbol{X}_1$ and $\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}$ are all invertible and, therefore, the operations are legitimate. Now consider the second summand in (17):

$$
\begin{aligned}
&\tilde{\boldsymbol{A}}\boldsymbol{\beta}_q\boldsymbol{\beta}_q^T\tilde{\boldsymbol{A}}^T = \begin{bmatrix} (\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{X}_q\boldsymbol{\beta}_q \\ (\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_q\boldsymbol{\beta}_q \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_q^T\boldsymbol{X}_2^T\boldsymbol{P}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{Z})^{-1} & \boldsymbol{\beta}_q^T\boldsymbol{X}_q^T\boldsymbol{Q}\boldsymbol{X}_p(\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_p)^{-1} \end{bmatrix} = \\
&\begin{bmatrix} (\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{X}_q\boldsymbol{\beta}_q\boldsymbol{\beta}_q^T\boldsymbol{X}_q^T\boldsymbol{P}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{Z})^{-1} & (\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{X}_q\boldsymbol{\beta}_q\boldsymbol{\beta}_q^T\boldsymbol{X}_q^T\boldsymbol{Q}\boldsymbol{X}_p(\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_p)^{-1} \\ (\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_q\boldsymbol{\beta}_q\boldsymbol{\beta}_q^T\boldsymbol{X}_q^T\boldsymbol{P}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{Z})^{-1} & (\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_q\boldsymbol{\beta}_q\boldsymbol{\beta}_q^T\boldsymbol{X}_q^T\boldsymbol{Q}\boldsymbol{X}_p(\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_p)^{-1} \end{bmatrix}
\end{aligned}
$$

Then the submatrix of (17) corresponding to the primary terms is

$$
\begin{aligned}
\mathrm{MSE}(\hat{\tilde{\boldsymbol{\beta}}}_p|\tilde{\boldsymbol{\beta}})_{pp} =& \sigma^2(\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_p)^{-1} + (\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_q\boldsymbol{\beta}_q\boldsymbol{\beta}_q^T\boldsymbol{X}_q^T\boldsymbol{Q}\boldsymbol{X}_p(\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_p)^{-1} \\
=& \sigma^2\tilde{\boldsymbol{M}}^{-1} + \tilde{\boldsymbol{M}}^{-1}\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_q\boldsymbol{\beta}_q\boldsymbol{\beta}_q^T\boldsymbol{X}_q^T\boldsymbol{Q}\boldsymbol{X}_p\tilde{\boldsymbol{M}},^{-1} \qquad (18)
\end{aligned}
$$

where $\tilde{\boldsymbol{M}} = \boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_p$.

As in the unblocked case, we first look at the determinant of the corresponding subma-

trix (18):

$$\begin{aligned}\det[\mathrm{MSE}(\hat{\tilde{\boldsymbol{\beta}}}_p|\tilde{\boldsymbol{\beta}})_{pp}] &= \det[\sigma^2\tilde{\boldsymbol{M}}^{-1} + \tilde{\boldsymbol{M}}^{-1}\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_q\boldsymbol{\beta}_q\boldsymbol{\beta}_q^T\boldsymbol{X}_q^T\boldsymbol{Q}\boldsymbol{X}_p\tilde{\boldsymbol{M}}^{-1}] = \\ &\sigma^{2p}\det[\tilde{\boldsymbol{M}}^{-1} + \tilde{\boldsymbol{M}}^{-1}\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_q\tilde{\boldsymbol{\beta}}_q\tilde{\boldsymbol{\beta}}_q^T\boldsymbol{X}_q^T\boldsymbol{Q}\boldsymbol{X}_p\tilde{\boldsymbol{M}}^{-1}] = \\ &\sigma^{2p}\det[\tilde{\boldsymbol{M}}^{-1}](1 + \tilde{\boldsymbol{\beta}}_q^T\boldsymbol{X}_q^T\boldsymbol{Q}\boldsymbol{X}_p^T\tilde{\boldsymbol{M}}^{-1}\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_q\tilde{\boldsymbol{\beta}}_q). \qquad (19)\end{aligned}$$

The q-dimensional random vector $\tilde{\boldsymbol{\beta}}_q$, as before, follows $\mathcal{N}(\boldsymbol{0}, \tau^2\boldsymbol{I}_q)$, so that this prior does not depend on the error variance $\sigma^2_\cdot$. Next, taking the expectation of the logarithm of (19) over the set prior distribution is completely identical to the derivations leading to (10). The $MSE(D)$-component then becomes:

$$\log(\det[\tilde{\boldsymbol{M}}^{-1}]) + \mathtt{E}_{\tilde{\boldsymbol{\beta}}_2}\log(1 + \tilde{\boldsymbol{\beta}}_2'\boldsymbol{X}_2'\boldsymbol{Q}\boldsymbol{X}_1\tilde{\boldsymbol{M}}^{-1}\boldsymbol{X}_1'\boldsymbol{Q}\boldsymbol{X}_2\tilde{\boldsymbol{\beta}}_2) \qquad (20)$$

and the resulting determinant-based compound criterion for a blocked experiments is

$$\begin{aligned}&\left[\left|(\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_p)^{-1}\right|^{1/p} F_{p,d_B;1-\alpha_{DP}}\right]^{\kappa_{DP}} \times \\ &\left[\left|\tilde{\boldsymbol{L}} + \frac{\boldsymbol{I}_q}{\tau^2}\right|^{-1/q} F_{q,d_B;1-\alpha_{LoF}}\right]^{\kappa_{LoF}} \times \qquad (21) \\ &\left[|\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_p|^{-1}\exp\left(\frac{1}{N}\sum_{i=1}^{N}\log(1 + \tilde{\boldsymbol{\beta}}_{2i}'\boldsymbol{X}_q^T\boldsymbol{Q}\boldsymbol{X}_p\tilde{\boldsymbol{M}}^{-1}\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_q\tilde{\boldsymbol{\beta}}_{2i})\right)\right]^{\kappa_{MSE}/p} \longrightarrow \min.\end{aligned}$$

Taking the expectation of the trace of (18) makes up the trace-based component criterion:

$$\begin{aligned}\mathtt{E}_{\beta_q}\mathrm{trace}[\mathrm{MSE}(\hat{\tilde{\boldsymbol{\beta}}}_p|\tilde{\boldsymbol{\beta}})_{pp}] &= \mathrm{trace}[\mathtt{E}_{\beta_q}\mathrm{MSE}(\hat{\tilde{\boldsymbol{\beta}}}_p|\tilde{\boldsymbol{\beta}})_{pp}] \\ &= \mathrm{trace}[\sigma^2\tilde{\boldsymbol{M}}_{pp}^{-1} + \mathtt{E}_{\beta_q}(\tilde{\boldsymbol{A}}\boldsymbol{\beta}_q\boldsymbol{\beta}_q^T\tilde{\boldsymbol{A}})_{pp}] \\ &= \sigma^2\mathrm{trace}[\tilde{\boldsymbol{M}}_{pp}^{-1} + \tau^2\{\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^T\}_{pp}] \\ &= \sigma^2[\mathrm{trace}(\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_p)^{-1} + \tau^2\mathrm{trace}\{\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^T\}_{pp}]. \qquad (22)\end{aligned}$$

The $MSE(LP)$-compound criterion for a blocked experiment is

$$\begin{aligned}&\left[\frac{1}{p}\mathrm{trace}(\boldsymbol{W}\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_p)^{-1}F_{1,d_B;1-\alpha_{LP}}\right]^{\kappa_{LP}} \times \\ &\left[\frac{1}{q}\mathrm{trace}\left(\tilde{\boldsymbol{L}} + \boldsymbol{I}_q/\tau^2\right)^{-1}F_{1,d_B;1-\alpha_{LoF}}\right]^{\kappa_{LoF}} \times \\ &\left[\frac{1}{p}\mathrm{trace}\{(\boldsymbol{X}_p^T\boldsymbol{Q}\boldsymbol{X}_p)^{-1} + \tau^2[\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^T]_{pp}\}\right]^{\kappa_{MSE}} \longrightarrow \min. \qquad (23)\end{aligned}$$

The probability levels α, weights κ hold the same meanings as in the unblocked case; and as was noted before, the number of pure error degrees of freedom d_B accounts for the comparisons between blocks.

4.3 Example. Case study

We shall consider an example of a real-life optimal design problem, and explore a range of solutions provided by the compound criteria constructed in previous sections. A company specialising in food supplements production for small animals were to conduct an experiment to figure out whether a slight decrease in the recommended dosages of three particular products taken together would make a significant impact on the resulting "performance", which is expressed in terms of some continuous response. The dosage range of interest for each supplement (experimental factor) is from 90% to 100% of the standard recommendation; it is desired that there would be 3 levels (i.e. taking the values of 90%, 95% and 100%). Carrying out the experiment with more than 3 levels was more complicated: measuring, for example, 92.5% of the recommended dosage was quite tricky.

The treatments were to be applied to $n = 36$ cages of animals (experimental units), which would be allocated in $b = 2$ equal sized blocks. The fullest response surface model, therefore, would contain all linear, quadratic and interaction terms: the primary model contains $p = 9$ terms. As it was suspected that increasing dosages beyond certain values might not be of an impact, that would be translated in a non-quadratic curvature of the fitted function – meaning that a presence of higher order terms would provide a better fit for the data, and it was desirable to accommodate for that possibility at the design stage. In the extended model We accounted for $q = 10$ potential terms ((linear-by-linear-by-linear, quadratic-by-linear and cubic) in the extended model, with the notations the same as before:

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta}_B + \boldsymbol{X}_p\boldsymbol{\beta}_p + \boldsymbol{X}_q\boldsymbol{\beta}_q + \boldsymbol{\varepsilon},$$

The design search was performed among a larger 5-level candidate set of points, but due to the form of the 2nd order polynomial primary model and the criteria used, the resulting designs were 3-level ones. The experimenters also wished to have at least two centre points in each block to ensure representation of the conditions thought a priori most likely to be best (with dosages of 95% for each supplement), i.e. 4 runs in total were fixed beforehand; this constraint was built directly into the search procedure; we also obtained designs without this restriction and evaluated the efficiency losses.

The experimenters preferred using the determinant-based criterion as the primary inferential interest is on the overall input of the model terms; we searched for exact optimal designs with respect to the compound MSE(DP)-criteria for blocked experiments (19). For each combination of weights we will consider two cases: $\tau^2 = 1$ and $\tau^2 = 1/q$; as for the number of Monte Carlo samples used to estimate the third criterion component, for$\tau^2 = 1$ we set $N = 500$, and for $\tau^2 = 1/q$ we set $N = 1000$ in order to have a sufficiently small relative estimation error ([cite the thesis?]). The search for each design was performed with 50 random starts. We considered three sets of weights: (1) with the

weight being equally distributed between the components, (2) a bit more weight (0.4) put on the DP-component, with the rest allocated to the lack-of-fit component, and (3) with half of the weight on the $MSE(D)$-component with the rest of it distributed evenly among the others.

Table 3 contains the summaries of the optimal designs; the two types of efficiencies are presented: with respect to the individual criteria with ("CP Efficiency") and without ("No CP Efficiency") the fixed two centre points. The former ones will be, obviously, larger, and the differences represent the magnitude of the loss by restricting the set of designs to be considered. The"Relative Efficiency" column reflects how well the given design performs with respect to the optimal one in terms of the same compound criterion, obtained without fixing the centre points, providing a "single-value" measure on the efficiency reduction.

Table 3: Case-study. Properties of MSE(D)-optimal blocked designs, with two centre points per block

	Criteria, $\tau^2 = 1$			**DoF**		**No CP Efficiency,%**			**CP Efficiency,%**			**Relative**
	DP	**LoF(DP)**	**MSE(D)**	**PE**	**LoF**	**DP**	**LoF(DP)**	**MSE(D)**	**DP**	**LoF(DP)**	**MSE(D)**	**Efficiency,%**
1	1/3	1/3	1/3	14	11	88.63	90.03	99.97	92.89	97.59	100.75	98.18
2	0.4	0.2	0.4	14	11	88.35	90.14	99.15	92.60	97.70	99.92	98.31
3	0.25	0.25	0.5	14	11	88.63	90.03	99.75	92.89	97.59	100.53	98.90
4	1	0	0	20	5	95.41	62.13	95.24	100.00	67.34	95.98	95.58
5	0	1	0	14	11	66.49	92.26	79.91	69.69	100.00	80.53	92.26
6	0	0	1	14	11	88.31	88.35	99.23	92.55	95.77	100.00	99.23
	Criteria, $\tau^2 = 1/q$			**DoF**		**No CP Efficiency,%**			**CP Efficiency,%**			**Relative**
	DP	**LoF(DP)**	**MSE(D)**	**PE**	**LoF**	**DP**	**LoF(DP)**	**MSE(D)**	**DP**	**LoF(DP)**	**MSE(D)**	**Efficiency,%**
1	1/3	1/3	1/3	18	7	92.52	95.13	95.86	96.97	98.00	96.55	94.95
2	0.4	0.2	0.4	18	7	94.19	92.19	96.54	98.72	94.98	97.24	95.34
3	0.25	0.25	0.5	17	8	92.44	93.70	97.05	96.88	96.53	97.75	95.72
4	1	0	0	20	5	95.41	90.63	95.66	100.00	93.37	96.35	95.41
5	0	1	0	22	3	76.11	97.07	77.10	79.77	100.00	77.65	97.07
6	0	0	1	14	11	88.37	90.37	99.29	92.62	93.10	100.00	99.29

The main feature observed is that in general individual efficiency values are quite large, both in the restricted or unrestricted cases. This might be attributed to the large number of available residual degrees of freedom (25), and this contributes to better compromises achievable between the three criterion components. The imbalance in the distribution of the residual degrees of freedom is not strong, however, the prevalence towards the pure error is quite consistent. In general, the performances are better for smaller τ^2 (except for the $MSE(D)$ part), as in this case the model disturbance effect is assumed to be quite small, and the compromise might be more feasible. Overall, relative efficiencies are quite good, losses among the first three designs (optimal with respect to the compound criteria) do not exceed 1.82% for τ^2 and 5.05% for $\tau^2 = 1/q$.

It is notable that designs #1 and #3 (for the larger τ^2) are the same. Its $MSE(D)$-value turned out to be better than of the design #6, which was constructed as being optimal with respect to this component. This design was chosen to be used when carrying out the experiment; it has been run successfully and useful conclusions were drawn from the obtained data. It can be found in Table 4 and on Figure 1, the points in each block have been ordered for the sake of easier perception, and, of course, they have to be and were randomised in each block before running the experiment. There

are only two centre points in each block, and replicates of other points are generally split between blocks (except for the $(-1, 1, 1)$ point which is duplicated in the first block only).

Table 4: Case-study. MSE(D)-optimal design #1 with two centre points, $\tau^2 = 1$

-1	-1	-1	-1	-1	-1
-1	-1	0	-1	-1	0
-1	-1	1	-1	-1	1
-1	0	-1	-1	0	1
-1	1	-1	-1	1	-1
-1	1	1	-1	1	0
-1	1	1	0	-1	-1
0	-1	1	0	-1	1
0	0	0	0	0	0
0	0	0	0	0	0
0	1	-1	0	1	1
1	-1	-1	1	-1	-1
1	-1	0	1	-1	0
1	-1	1	1	-1	1
1	0	1	1	0	-1
1	1	-1	1	0	1
1	1	0	1	1	-1
1	1	1	1	1	1

As for the time costs, on average an optimal design was found in $13 - 15$ hours, which was acceptable in this particular case. Sometimes, however, it took up to $20 - 24$ hours, so some extra time allowance should be accounted for when using these criteria and this search algorithm and/or the extensive sampling might be replaced by a less demanding alternative.

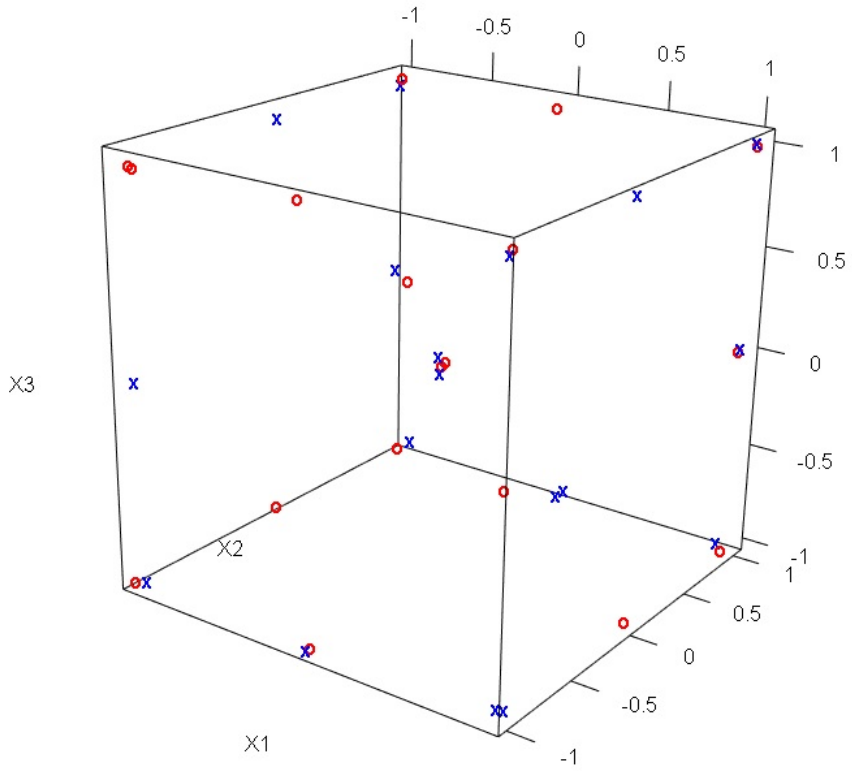

Figure 1: MSE(D)-optimal design #1. Colours (blue and red) and symbols ('x' and 'o') serve as block indicators.

[OE: I'm not sure we need to include the following: looking at what would have happened if we could afford 5 levels.]

Although the primary choice of the number of levels of each factor is 3, the number of experimental runs would allow for 5 levels so that all of the potential terms can be estimated. So we considered one more compound criterion which has a different lack-of-fit component: DP_S– optimality for potential terms in the full (third-order polynomial) model, arising from the D_S-optimality that was suggested by Atkinson *et al.* (2007) (page 360) in the context of model discrimination:

$$\left[\left|\tilde{\boldsymbol{M}}\right|^{-1/p} \mathrm{F}(p, d, \alpha_{DP})\right]^{\kappa_{DP}} \times \left[|\tilde{\boldsymbol{L}}|^{-1/q} F_{q,d_B;1-\alpha_{LoF}}\right]^{\kappa_{LoF}} \times$$
$$\left[|\tilde{\boldsymbol{M}}|^{-1} \exp\left(\frac{1}{N}\sum_{i=1}^{N} \log(1 + \tilde{\boldsymbol{\beta}}_{2i}' \boldsymbol{X}_q^T \boldsymbol{Q} \boldsymbol{X}_1 \tilde{\boldsymbol{M}}^{-1} \boldsymbol{X}_1' \boldsymbol{Q} \boldsymbol{X}_q \tilde{\boldsymbol{\beta}}_{2i})\right)\right]^{\kappa_{MSE}/p}. \quad (24)$$

This criterion further on will be referred to as the "full" criterion, and we shall explore how this would affect the resulting designs.

The summary of the corresponding optimal designs is given in Table 5: the 'new' lack-of-fit component is denoted by DP*s. The notions of 'No CP' and 'CP' efficiencies are the same as previously. For each design we calculated its $LoF(DP)$-value, so that we could assess how they would perform in terms of the criterion (6). The performances of

the optimal designs are summarised in Table 5. In order to illustrate general tendencies in the designs' appearances, designs optimal with respect to the criterion with equal weights, for two values of τ^2 are provided in Table 6.

Table 5: Case-study. Properties of "Full" MSE(D)-optimal blocked designs, with two centre points per block

	Criteria, $\tau^2 = 1$			DoF		No CP Efficiency,%				CP Efficiency,%				Relative
	DP	DP*s	MSE(D)	PE	LoF	DP	DP*s	LoF(DP)	MSE(D)	DP	DP*s	LoF(DP)	MSE(D)	Efficiency,%
1	1/3	1/3	1/3	14	11	80.58	69.38	87.05	91.74	84.46	79.90	94.36	92.46	94.42
2	0.4	0.2	0.4	14	11	83.90	61.18	85.36	57.47	87.93	70.46	92.52	57.92	95.83
3	0.25	0.25	0.5	13	12	79.93	68.24	87.06	94.05	83.77	78.59	94.37	94.79	96.32
4	1	0	0	20	5	95.41	0.00	62.13	95.24	100.00	0.00	67.34	95.98	95.41
5	0	1	0	14	11	62.01	86.83	92.98	74.95	64.99	100.00	100.78	75.53	86.83
6	0	0	1	14	11	88.31	0.00	88.35	99.23	92.55	0.00	95.77	100.00	99.23

	Criteria, $\tau^2 = 1/q$			DoF		No CP Efficiency,%				CP Efficiency,%				Relative
	DP	DP*s	MSE(D)	PE	LoF	DP	DP*s	LoF(DP)	MSE(D)	DP	DP*s	LoF(DP)	MSE(D)	Efficiency,%
1	1/3	1/3	1/3	14	11	80.14	69.01	87.70	91.41	83.99	81.43	90.34	92.07	93.97
2	0.4	0.2	0.4	14	11	83.76	61.29	88.28	94.41	87.79	72.31	90.95	95.09	96.04
3	0.25	0.25	0.5	12	13	79.66	64.43	84.44	94.84	83.49	76.02	86.99	95.52	95.20
4	1	0	0	20	5	95.41	0.00	90.63	95.66	100.00	0.00	93.37	96.35	95.41
5	0	1	0	14	11	59.81	84.75	86.15	72.05	62.68	100.00	88.75	72.57	84.75
6	0	0	1	14	11	88.73	0.00	90.57	99.29	92.99	0.00	93.30	100.00	99.29

Table 6: Case-study. Designs #1 from Table 5, $\tau^2 = 1$ (left) and $\tau^2 = 1/q$ (right)

-1	-1	0	-1	-1	-1		-1	-1	-1	-1	-1	-1
-1	-1	1	-1	-1	0		-1	-1	0.5	-1	-1	0.5
-1	0	-1	-1	-1	1		-1	-0.5	1	-1	-0.5	1
-1	0.5	1	-1	0.5	1		-1	0	-0.5	-1	1	-1
-1	1	-1	-1	1	-1		-1	1	-1	-1	1	1
-1	1	0.5	-1	1	0.5		-1	1	1	-0.5	-1	-0.5
-0.5	1	1	-0.5	-0.5	1		-0.5	1	0	-0.5	0.5	-1
0	-1	-1	-0.5	1	-0.5		0	-1	1	0	-1	1
0	0	0	0	-1	-1		0	0	0	0	0	0
0	0	0	0	0	0		0	0	0	0	0	0
0.5	-1	1	0	0	0		0	1	1	0	1	1
0.5	1	-1	0.5	-1	1		0.5	-0.5	-1	0.5	1	-0.5
1	-1	-1	0.5	1	-1		1	-1	-1	1	-1	-1
1	-1	0.5	1	-1	-1		1	-1	0	**1**	**-1**	**1**
1	-0.5	1	1	-1	0.5		1	0.5	1	**1**	**-1**	**1**
1	0.5	-1	1	-0.5	-0.5		**1**	**1**	**-1**	1	0	-0.5
1	1	-0.5	1	0.5	-1		**1**	**1**	**-1**	1	0.5	1
1	1	1	1	1	1		1	1	0.5	1	1	0.5

In case of $\tau^2 = 1$ all pure error degrees of freedom (except for the 2 coming from the replicated centre points) occur from 12 points duplicated in different blocks; in case of $\tau^2 = 1/q$ (i.e. $\tau^2 = 0.1$) two 'corner' points are replicated within the same block (they are highlighted in Table 6). Quite a few experimental units would receive an 'intermediate' ± 0.5 dosage of at least one product, and as this did not comply with the demands of the experimenters, the choice was still made in favour of the three-level design in Table 4.

5 Multistratum Experiments

In a large number of industrial, engineering and laboratory-based experiments, either the nature of the study or certain restrictions result in the necessity of considering a multi-level structure of experimental units. For example, a chemical process consisting of applying treatments to the material batches of different sizes at each stage; or one

or several experimental factors' values can be changed only once per a certain amount of runs whereas values of other factors are varied between runs. Therefore, different factors are applied at different levels, and randomisation is performed at each level, thus the whole process of allocating treatments to experimental units should be amended accordingly (Mead *et al.*, 2012).

In this work we go on to consider experimental framework comprising such a hierarchical structure of experimental units and treatments, which is referred to as multistratum experiment, and each stratum is defined as a level in the unit structure. Units are grouped into whole-plots, each of them divided into sub-plots, which contain a certain number of sub-sub-plots, and so on up to the smallest units — runs of the experiment – [Figure with Hasse diagram?] . In the general case of two strata, we deal with what is called a "split-plot" experiment, in case of three strata a "split-split-plot" experiment.

Our aim is to adapt the MSE-based criteria derived before to the factorial experiments with units distributed in several strata. The most important aspects are (1) estimating the error components at every level and (2) formulating the compliant criteria and presenting a suitable implementation strategy for the optimal design search.

Here we are considering experiments with s strata in total, stratum i being nested within units of the stratum $i-1$, and stratum 0 will be seen as the whole experiment (Trinca and Gilmour (2015)). The number of units in stratum i within every unit in stratum $i-1$ is denoted by n_i, such that $m_j = \prod_{i=1}^{j} n_i$ is the number of units in stratum j and, therefore, $n = m_s = \prod_{i=1}^{s} n_i$ is the total number of runs. As randomisation occurs at each stratum, the overall polynomial model accounting for the hierarchical error structure and correlated observations can be written as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \sum_{i=1}^{s} \boldsymbol{Z}_i \boldsymbol{\varepsilon}_i, \tag{25}$$

where $\boldsymbol{Y}$ represents the n-dimensional vector of responses, $\boldsymbol{X}$ stands for the $n \times p$ model matrix and $\boldsymbol{\beta}$ for the p-dimensional vector of corresponding model coefficients. Each row of the $n \times m_i$ matrix $\boldsymbol{Z}_i$ corresponds to a single experimental run and indicates the unit in stratum i containing this run. $\boldsymbol{\varepsilon}_i$ is a vector of random effects occurring due to the randomisation at level i, and these effects are assumed to be independent and identically distributed around zero mean and variance σ_i^2. Generalised Least Squared (GLS) estimators of $\boldsymbol{\beta}$ are calculated as

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y}. \tag{26}$$

An extensive amount of research has been conducted on the design of and analysis of data from split-plot and split-split-plot experiments, with one of the first works by Letsinger *et al.* (1996), emphasising the necessity of adapting the design and analysis strategy with respect to the more complicated error structure. In the most general case of non-orthogonal unit structures Residual Maximum Likelihood (REML) methodology

has been acknowledged to be the most sensible approach to estimating the variance components by maximising the part of the likelihood function corresponding to the "random" part of the model; the details are provided, for example, in McCulloch and Searle (2001).

In some case, though, REML tends to underestimate the variance components in the higher stratum, for example Goos *et al.* (2006) considered some split-plot experiments with true whole-plot variances being non-negligibly far from zero, for which REML, however, provided zero estimates. A possible alternative of following Bayesian strategy (Gilmour and Goos, 2009), however, this requires a careful choice of the prior which is often not possible, especially in the case of more than two strata.

In the context of model uncertainty, it would be appropriate to consider instead the direct relationship between the response and the set of treatments, similar to (2):

$$\boldsymbol{Y} = \boldsymbol{X_t}\boldsymbol{\mu_t} + \sum_{i=1}^{s} \boldsymbol{Z}_i\boldsymbol{\varepsilon}_i. \tag{27}$$

Each column of $\boldsymbol{X_t}$ corresponds to a treatment defined as a combination of the experimental factors; $\boldsymbol{Z}_i$ is the indicator matrix of random effects at stratum i.

Gilmour *et al.* (2017b) introduced the notion of "Pure Error REML", where random effects are estimated from the treatment model (27), and then for the sake of making inference regarding parameters of the response model (25), they are substituted in the GLS estimators of the fixed effects (26). It is also argued that some appropriate corrections are to be adapted and applied to the obtained estimates, i.e. the ones suggested by Kenward and Roger (1997).

A careful definition of "pure error" is necessary when considering the treatment and response surface models, especially in the presence of nested random block effects, as discussed by Gilmour and Trinca (2000). Treatment-block additivity assumption (Draper and Smith (1998)) would allow for estimating pure error not just from the within-block replicates, but also using inter-block information; (Gilmour and Trinca (2000)) argue for preserving the treatment-unit additivity as well, for the consistency with the unblocked cases – so that between unit variability is measured regardless of the treatments applied. Based on the interrelation of the polynomial and full treatment models, and the desirability of being able to provide the possibility for testing for the lack-of-fit together with obtaining robust estimates of the variance components, the approach of using pure error is adopted, Gilmour *et al.* (2017a) explored several approaches to estimating the variance components.

The first approach is stratum-by-stratum data analysis, where each randomisation level is considered separately, and at each stratum i units are treated as runs aggregated in m_{i-1} blocks with fixed effects. The lower stratum variance is estimated from within blocks (using intra-block residual mean square), and the higher stratum (inter-block)

variance from the difference between the intra- and inter-block residual mean squares scaled by the number of runs per block (Hinkelmann and Kempthorne, 2005). However, it would be desirable to make use of inter-block replicates as well, especially in the common cases of relatively small experiments; Yates' procedure, first suggested by Yates (1939) and described in detail by Hinkelmann and Kempthorne (2005), provides more degrees of freedom for the variance estimate in the higher stratum – and this is the approach we are adopting in this work.

In the case of two levels of randomisation, variance components are estimated in two variance decomposition steps:

1. Total SS = SS(Blocks) + SS(Treatments|Blocks) + SS(Residual)
   From fitting this full treatment-after-blocks model the residual mean square S^2 is taken as the estimate of the intra-block variance. SS(Residual) is then substituted in the following partition.

2. Total SS = SS(Treatments) + SS(Blocks|Treatments) + SS(Residual)
   Therefore, the sums of squares for the blocks after the treatment effects have been accounted for is obtained and the corresponding mean square

   $$S_b^2 = \text{SS(Blocks|Treatments)}/\nu_b$$

   is then used to get an estimate of the inter-block variance: $\hat{\sigma}_b^2 = \frac{\nu_b(S_b^2 - S^2)}{d}$, where pure error degrees of freedom for inter-block variance is $d = n - \text{trace}[\boldsymbol{Z}'\boldsymbol{X_t}(\boldsymbol{X_t}'\boldsymbol{X_t})^{-1}\boldsymbol{X_t}'\boldsymbol{Z}]$, $\nu_b = \text{rank}([\boldsymbol{X_t}\boldsymbol{Z}]) - \text{rank}(\boldsymbol{X_t})$, and $\boldsymbol{X_t}$ and $\boldsymbol{Z}$ are as in (27).

Allocation of the degrees of freedom in practice, following Yates' procedure, is provided in the Appendix A.2.

Gilmour *et al.* (2017a) show that in addition to the treatment-unit additivity and randomisation the assumption of the experimental units being a random sample from an infinite population is necessary for the inter-block variance estimate to be unbiased. In this particular context, when the normality and independence of responses in (25) is a standard assumption, although a potential presence of contamination effects in the fixed part of the model is also accounted for, Yates' approach seems to be the most appropriate technique. Relying on the distribution assumption to the extent that makes REML the most sensible approach means that the fixed part of the fitted model is assumed to be absolutely true, that is the parameters of the population distribution are known, which does not comply with the model uncertainty framework.

In the presence of potential model disturbance which is expressed as q additional polynomial terms for the purposes of optimal planning, the overall model for a multistratum experiment is similar to (5):

$$\boldsymbol{Y} = \boldsymbol{X}_p\boldsymbol{\beta}_p + \boldsymbol{X}_q\boldsymbol{\beta}_q + \sum_{i=1}^{s} \boldsymbol{Z}_i\boldsymbol{\varepsilon}_i. \tag{28}$$

The design construction strategy has been chosen to comply with the error estimation and the whole framework, and will follow the stratum-by-stratum algorithm approach developed by Trinca and Gilmour (2015). Starting with the higher stratum, it is an iterative procedure that at each step treats the higher stratum units as fixed blocks and generates a candidate set of treatments applied at the current stratum. The model matrix then comprises terms inherited from all higher strata, the current stratum's polynomial terms and interactions between these two groups of terms; the current stratum search is carried out using the same algorithm, e.g. point exchange. Treating the previous strata as fixed blocked effects means that the optimality criteria used are exactly the ones that were introduced for unblocked (Section 3) and blocked (Section 4) experiments. The detailed step-by-step procedure is outlined in Appendix A.3, and an example of a multistartum experimental framework and optimal design search is considered next.

5.1 Split-split-plot optimal design

As an illustrative example, we consider a multistratum experimental framework with three strata in total, and 2, 1 and 2 factors applied in each of them respectively. 48 experimental unites are set up in 12 whole-plots, each of them containing 2 sub-plots with 2 runs per sub-plot. The fitted full quadratic polynomial model in this case contains 20 parameters: 5 in the first stratum, 4 in the second and 11 in the third. As for the potential terms, there are 30 of them: just 2 and 5 are in the whole- and sub-plots, the other 23 are the third-order terms formed from the factors in the sub-sub-plots and their interactions with the terms from the higher strata.

The summaries of $MSE(DP)$- and $MSE(LP)$-optimal designs' performances and the distributions of the degrees of freedom are given in Tables 7 and 8 respectively. In the first two strata some of the degrees of freedom are allocated to the inter-whole-plot ("Inter-WP") and inter-sub-plot ("Inter-SP") information components.

Even though there are not many residual degrees of freedom available at each stratum, the pattern of the imbalance between the pure error and lack-of-fit components can be observed: lower values of the variance scaling parameter τ^2 result in more replicate. In the case of $MSE(LP)$-optimal designs this pattern seems to be coherent with the weight put on the LP- and $LoF(LP)$-components.

In general, similar to the examples seen before, searching for compound optimality results in designs performing decently in terms of DP-, LP- and lack-of-fit components, with the efficiencies dropping when no weight is allocated to the corresponding part of the criterion and/or with the value of τ^2 increasing. Increased variance scaling parameter - meaning larger prior scale of the fitted model disturbance - leads to generally higher $MSE(D)$-efficiency of the optimal designs, that is protecting the quality of the primary model parameters' estimates from the potential deviation.

Table 7: Summary of MSE(DP)-optimal split-split-plot designs

	Criteria, $\tau^2 = 1/q$			DF (s=1)			DF (s=2)			DF(s=3)		Efficiencies,%		
	DP	LoF(DP)	MSE(D)	PE	Inter-WP	LoF	PE	Inter-SP	LoF	PE	LoF	DP	LoF(DP)	MSE(D)
1	1	0	0	3	0	3	5	1	2	10	3	100.00	91.44	85.32
2	0	1	0	2	1	3	3	5	0	13	0	5.99	100.00	5.06
3	0	0	1	0	3	3	0	3	5	0	13	0.00	0.00	100.00
4	1/3	1/3	1/3	3	0	3	3	3	2	9	4	92.85	87.17	83.81
5	0.5	0.5	0	2	1	3	6	1	1	11	2	84.77	96.02	70.81
6	0.5	0	0.5	3	0	3	4	1	3	8	5	93.22	83.11	90.42
7	0	0.5	0.5	3	0	3	2	4	2	9	4	85.98	90.59	79.33
8	0.5	0.25	0.25	3	0	3	4	2	2	9	4	90.88	91.03	83.39

	Criteria, $\tau^2 = \sqrt{1/q}$			DF (s=1)			DF (s=2)			DF(s=3)		Efficiencies,%		
	DP	LoF(DP)	MSE(D)	PE	Inter-WP	LoF	PE	Inter-SP	LoF	PE	LoF	DP	LoF(DP)	MSE(D)
1	1	0	0	3	0	3	5	1	2	10	3	100.00	92.11	84.07
2	0	1	0	1	2	3	2	5	1	10	3	30.24	100.00	28.05
3	0	0	1	0	3	3	0	3	5	0	13	0.00	0.00	100.00
4	1/3	1/3	1/3	2	1	3	5	1	2	9	4	86.39	96.27	78.05
5	0.5	0.5	0	2	1	3	3	3	2	9	4	83.93	100.76	76.45
6	0.5	0	0.5	3	0	3	5	0	3	9	4	96.90	90.06	86.09
7	0	0.5	0.5	3	0	3	2	3	3	8	5	77.78	97.70	78.21
8	0.5	0.25	0.25	3	0	3	2	4	2	9	4	88.17	95.22	79.23

	Criteria, $\tau^2 = 1$			DF (s=1)			DF (s=2)			DF(s=3)		Efficiencies,%		
	DP	LoF(DP)	MSE(D)	PE	Inter-WP	LoF	PE	Inter-SP	LoF	PE	LoF	DP	LoF(DP)	MSE(D)
1	1	0	0	3	0	3	5	1	2	10	3	100.00	69.23	89.76
2	0	1	0	2	2	2	1	4	3	7	6	46.43	100.00	51.81
3	0	0	1	1	2	3	0	3	5	0	13	0.00	0.00	100.00
4	1/3	1/3	1/3	3	0	3	5	0	3	8	5	83.38	84.60	88.46
5	0.5	0.5	0	1	2	3	3	2	3	8	5	74.06	87.14	78.63
6	0.5	0	0.5	3	0	3	5	0	3	9	4	98.09	74.68	93.25
7	0	0.5	0.5	3	0	3	3	2	3	6	7	81.25	94.80	93.70
8	0.5	0.25	0.25	3	0	3	2	3	3	8	5	83.00	87.46	87.02

Most of the compound-optimal designs can provide a satisfying compromise among the three objectives; however, $MSE(D)-/MSE(L)-$ efficiency values are more responsive to the allocation of weights; so once an MSE-based component is of importance, this tendency certainly needs to be taken into account regardless of what criterion and variance scaling parameter is used.

An interesting observation worth noting is the $MSE(DP)$-optimal design #5 (for $\tau^2 = \sqrt{1/q}$) that was obtained as DP- and $LoF(DP)$-efficient turned out to be slightly more $LoF(DP)$-efficient (100.76%) rather than the one that was searched for as $LoF(DP)$-optimal (#2); besides, it performs much better in terms of the other criterion components, i.e. it is more than 70% efficient w.r.t. each of them. This is not something that would be normally expected, but the original design search procedure is expected to find near-optimum designs, especially when the expectation in the MSE(DP)-criterion is approximated – and that holds for both unblocked, blocked and multistratum frameworks.

6 Discussion

- *Conclusions*
- *Focus on future work(?) and/or alternatives*

Table 8: Example 2. Properties of MSE(L)-optimal designs

	Criteria, $\tau^2 = 1/q$			DF (s=1)			DF (s=2)			DF(s=3)		Efficiencies,%		
	LP	**LoF(LP)**	**MSE(L)**	**PE**	**Inter-WP**	**LoF**	**PE**	**Inter-SP**	**LoF**	**PE**	**LoF**	**LP**	**LoF(LP)**	**MSE(L)**
1	1	0	0	2	1	3	3	3	2	10	3	100.00	84.14	53.61
2	0	1	0	1	2	3	3	5	0	13	0	0.00	100.00	0.00
3	0	0	1	0	3	3	0	1	7	0	13	0.00	0.00	100.00
4	1/3	1/3	1/3	2	1	3	4	2	2	10	3	88.77	84.09	59.95
5	0.5	0.5	0	3	0	3	2	5	1	12	1	94.41	95.38	48.74
6	0.5	0	0.5	3	0	3	2	3	3	9	4	89.77	76.14	67.58
7	0	0.5	0.5	3	0	3	3	2	3	11	2	81.78	91.51	59.73
8	0.5	0.25	0.25	3	0	3	2	4	2	10	3	95.82	83.35	57.59

	Criteria, $\tau^2 = \sqrt{1/q}$			DF (s=1)			DF (s=2)			DF(s=3)		Efficiencies,%		
	LP	**LoF(LP)**	**MSE(L)**	**PE**	**Inter-WP**	**LoF**	**PE**	**Inter-SP**	**LoF**	**PE**	**LoF**	**LP**	**LoF(LP)**	**MSE(L)**
1	1	0	0	2	1	3	3	3	2	10	3	100.00	88.95	31.38
2	0	1	0	0	3	3	2	5	1	13	0	0.00	100.00	0.00
3	0	0	1	0	3	3	0	0	8	0	13	0.00	0.00	100.00
4	1/3	1/3	1/3	2	1	3	2	4	2	10	3	85.15	88.87	39.06
5	0.5	0.5	0	2	1	3	2	5	1	11	2	91.36	93.49	31.90
6	0.5	0	0.5	3	0	3	3	2	3	9	4	93.18	82.34	41.75
7	0	0.5	0.5	1	2	3	1	4	3	9	4	57.49	81.98	53.67
8	0.5	0.25	0.25	2	1	3	3	3	2	9	4	94.10	83.43	38.75

	Criteria, $\tau^2 = 1$			DF (s=1)			DF (s=2)			DF(s=3)		Efficiencies,%		
	LP	**LoF(LP)**	**MSE(L)**	**PE**	**Inter-WP**	**LoF**	**PE**	**Inter-SP**	**LoF**	**PE**	**LoF**	**LP**	**LoF(LP)**	**MSE(L)**
1	1	0	0	2	1	3	3	3	2	10	3	100.00	91.95	18.79
2	0	1	0	3	1	2	4	1	3	13	0	0.00	100.00	0.00
3	0	0	1	0	3	3	0	0	8	1	12	0.00	0.00	100.00
4	1/3	1/3	1/3	2	1	3	0	5	3	9	4	76.38	86.50	29.72
5	0.5	0.5	0	2	1	3	2	4	2	11	2	91.18	95.72	17.55
6	0.5	0	0.5	2	1	3	0	5	3	8	5	72.48	79.24	34.81
7	0	0.5	0.5	0	3	3	1	4	3	8	5	37.45	78.53	49.34
8	0.5	0.25	0.25	3	1	2	3	2	3	10	3	93.63	91.72	23.68

7 Acknowledgements

A Appendices

A.1 Appendix I

The DP-optimal design is given in Table 9 in Appendix **??**; it performs well in terms of the $MSE(D)$-component for both values of τ^2, however, its $LoF-DP$-efficiency drops by roughly half when the scaling parameter goes from $1/q$ to 1; also, $LoF-DP$-optimal designs provide the lowest DP-efficiency values (around 40%) for any value of τ^2.

Table 9: DP-optimal design

1	-1	-1	-1	-1	-1	21	0	-1	-1	1	0
2	-1	-1	-1	-1	-1	22	0	0	0	-1	1
3	-1	-1	0	1	1	23	0	0	0	-1	1
4	-1	-1	1	-1	1	24	0	1	-1	0	-1
5	-1	-1	1	-1	1	25	1	-1	-1	-1	1
6	-1	-1	1	1	-1	26	1	-1	-1	-1	1
7	-1	-1	1	1	-1	27	1	-1	-1	1	-1
8	-1	0	-1	0	1	28	1	-1	-1	1	-1
9	-1	0	-1	0	1	29	1	-1	1	-1	-1
10	-1	1	-1	-1	1	30	1	-1	1	-1	-1
11	-1	1	-1	-1	1	31	1	-1	1	1	1
12	-1	1	-1	1	-1	32	1	-1	1	1	1
13	-1	1	-1	1	-1	33	1	0	-1	-1	0
14	-1	1	0	0	0	34	1	1	-1	-1	-1
15	-1	1	0	0	0	35	1	1	-1	1	1
16	-1	1	1	-1	-1	36	1	1	-1	1	1
17	-1	1	1	-1	-1	37	1	1	1	-1	1
18	-1	1	1	1	1	38	1	1	1	-1	1
19	-1	1	1	1	1	39	1	1	1	1	-1
20	0	-1	-1	1	0	40	1	1	1	1	-1

A.2 Appendix II

Allocation of degrees of freedom

The distribution of the degrees of freedom among different components at each stratum of the resulting design needs to be evaluated in accordance with the variance estimation procedure (as described in Section **??**), so that both intra- and inter-block information is taken into account.

We shall consider here the case of a split-split-plot experiment (i.e. with 3 strata), but this strategy is straightforwardly extended to the cases with any number of randomisation levels. At the first 2 strata, where experimental units are to be further expanded, and treatments applied at these levels are to be replicated for all sub-units and runs within then the available degrees of freedom (after fitting the model at this current level) are split between three components: pure error, inter-plot (inter-whole-plot and inter-sub-plot for the 1st and 2nd strata respectively) and lack-of-fit. In the lowest stratum there are, as in the unblocked cases and experiments with fixed block effects, only pure error and lack-of-fit components.

Starting with presenting the way of calculating the number of pure error degrees of freedom at each level, we then proceed to the details of the inter-plot and lack-of-fit degrees of freedom evaluation accompanied by the corresponding R code. In the end of this section we will consider an illustrative example of a design and see how degrees of freedom are allocated.

If we consider an experiment with n_1 units in the first stratum, with each of them having n_2 units of stratum 2, and with n_3 runs nested within each of those, there are $n = n_1 \times n_2 \times n_3$ runs in total. The notation to be used is as follows:

- $\boldsymbol{y}$ – an n-dimensional dummy response vector (any randomly generated vector), which is used for fitting a linear model;
- $\boldsymbol{B_1}$ – an n-dimensional vector, such that its i-th element is the label of the first stratum unit that contains the i-th run. Similarly, $\boldsymbol{B_2}$ contains labels of the second stratum units;
- $\boldsymbol{T_W}$ – a vector of labels for treatments applied at the first stratum (whole-plots); $\boldsymbol{T_S}$ – a vector of labels for treatments applied at the second stratum units (sub-plots); this includes treatments in the higher stratum. And, finally, $\boldsymbol{T_{SS}}$ contains labels for all treatments applied at the lowest stratum, i.e. at individual runs;
- $\boldsymbol{Xw.m}$, $\boldsymbol{Xs.m}$ and $\boldsymbol{Xss.m}$ are the matrices containing fixed model terms that are to be fitted at the 1st, 2nd and 3rd strata respectively.

If we denote by p_i the number of the fixed model terms fitted at the i-th stratum, then the number of available degrees of freedom at each level is: $n_1 - 1 - p_1$ for the 1st, $n_1 \times (n_2 - 1) - p_2$ – for the 2nd, and $n_1 \times n_2 \times (n_3 - 1) - p_3$ – for the 3rd stratum.

After a design has been obtained, and all the treatment labels, block indicators and model matrices are in place, fitting the following linear models and studying their summary (i.e. ANOVA) will provide the numbers of degrees of freedom corresponding to every component.

We start with fitting the full treatment model schematically presented in (29): the effects of treatments applied to all strata, followed by the terms containing strata indicators, '$\boldsymbol{B_i}$'. In the resulting summary we then will see the distribution of pure error degrees of freedom: in the row corresponding to '$\boldsymbol{B_1}$' for the whole-plots (1st stratum), and in the row corresponding to '$\boldsymbol{B_2}$' for the sub-plots (2nd stratum). The 'Residual' number of degrees of freedom is equal to the number of pure error degrees of freedom for the 3rd stratum.

$$\mathrm{lm}(\boldsymbol{y} \sim \boldsymbol{T_{SS}} + \boldsymbol{B_1} + \boldsymbol{B_2}). \tag{29}$$

The order of terms in the expression above is crucial: this is a block-after-treatment model fitting; and the pure error degrees of freedom for lower-level plots occur not only when they are replicated within the same higher-level plot, but also when some pairs of

treatments are split between two (or more) higher-level plots. An example of a design is given at the end of this section together with a detailed description of how each pure error degree of freedom is obtained.

From the summary of model (29) we automatically obtain the number of lack-of-fit degrees of freedom (LoF) for the lowest stratum, as $n_1 \times n_2 \times (n_3 - 1) - p_3 =$'Residual' P.E. + LoF.

Now we need to evaluate the number of degrees of freedom for estimating the inter-whole-plot/inter-sub-plot and lack-of-fit components for the first two strata. For this purpose, at each of the higher strata, we fit a linear model comprising polynomial terms (formed of the factors applied at the current and all higher strata), treatment labels, and current and all higher strata unit indicators. By doing so and first projecting the vector of (dummy) responses to the subspaces generated by the parametric model's terms and by the treatments, we account for the degrees of freedom required for evaluating the model parameters' estimates and for the lack-of-fit degrees of freedom. The number of degrees of freedom left for the components corresponding to the current stratum indicators is the sum of the inter-whole-plot (inter-sub-plot) and pure error degrees of freedom.

$$\mathrm{lm}(\boldsymbol{y} \sim \boldsymbol{X w.m} + \boldsymbol{T_W} + \boldsymbol{B_1}), \tag{30}$$

$$\mathrm{lm}(\boldsymbol{y} \sim \boldsymbol{X w.m} + \boldsymbol{X s.m} + \boldsymbol{T_S} + \boldsymbol{B_1} + \boldsymbol{B_2}). \tag{31}$$

Therefore, the summary of the model in (30) provides the sum of inter-whole-plot and pure error degrees of freedom under the '$\boldsymbol{B_1}$' component; and the '$\boldsymbol{B_2}$' component of the the summary of (31) gives the same sum for the sub-plot level. Hence, by substituting the number of pure error degrees of freedom, we obtain the inter-whole-plot/inter-sub-plot and then, from the known total numbers, – the lack-of-fit degrees of freedom for the first and seconds strata.

In Table 10 below an example of a split-split-plot design is provided, with 12 whole-plots, each containing 2 sub-plots of size 2, i.e. 48 runs in total.

We see that whole-plots I and II, and IV and V are identical, but with the sub-plot and sub-sub-plot treatments being different, so each such replicate provides 1 whole-plot (WP), 1 sub-plot(SP) and 2 sub-sub-plot (SSP) pure error degrees of freedom. Whole-plots III, VI, VII and XI contain replicated sub-plots, although the sub-sub-plot treatments are not replicated within each sub-plot. Therefore, each of them provides 1 SP and 1 SSP pure error degree of freedom.

Finally, whole-plots VIII, IX and X contain 1 replicated point (SSP treatment) within one of the sub-plots, which provides another 3 SSP pure error degrees of freedom. In total, for three strata, there are 2, 6 and 11 pure error degrees of freedom respectively.

Table 10: Split-split-plot design

I	-1	-1	-1	-1	-1	VII	0	0	-1	1	0
	-1	-1	-1	1	1		0	0	-1	0	0
	-1	-1	1	-1	1		0	0	-1	1	0
	-1	-1	1	1	1		0	0	-1	0	0
II	-1	-1	-1	-1	-1	VIII	0	1	0	1	0
	-1	-1	-1	1	1		0	1	0	1	0
	-1	-1	1	-1	1		0	1	-1	1	1
	-1	-1	1	1	1		0	1	-1	-1	0
III	-1	0	1	0	-1	IX	1	-1	1	1	1
	-1	0	1	1	0		1	-1	1	-1	-1
	-1	0	1	0	-1		1	-1	-1	0	1
	-1	0	1	1	0		1	-1	-1	0	1
IV	-1	1	0	-1	1	X	1	-1	1	-1	1
	-1	1	0	1	-1		1	-1	1	1	-1
	-1	1	1	-1	-1		1	-1	-1	0	0
	-1	1	1	0	1		1	-1	-1	0	0
V	-1	1	0	-1	1	XI	1	0	-1	-1	-1
	-1	1	0	1	-1		1	0	-1	1	0
	-1	1	1	-1	-1		1	0	-1	-1	-1
	-1	1	1	0	1		1	0	-1	1	0
VI	0	-1	-1	-1	0	XII	1	1	1	-1	0
	0	-1	-1	1	-1		1	1	1	1	-1
	0	-1	-1	-1	0		1	1	-1	0	-1
	0	-1	-1	1	-1		1	1	-1	-1	1

A.3 Appendix III

Then the optimal multistratum design search is implemented following the steps below:

1. Starting from the first stratum, if there are any factors applied at this level, a candidate set of treatments is formed together with the fitted model matrix comprising the primary terms and the matrix of potential terms. The optimal unblocked design X_1 is then obtained using the usual point-exchange algorithm, by minimising (**??**) or (**??**). The labels of the treatments applied at the current stratum are saved at this stage as well in order to calculate the number of pure error degrees of freedom at the lower strata.

   If there are no factors applied at the first stratum, we move to the second one, and conduct the optimal design search for a blocked experiment, where the number of blocks is equal to the number of units in the first stratum n_1, with n_2 runs per block; in this case criterion function (**??**) or (**??**) is used, and number of pure error degrees of freedom d_B is calculated according to the usual blocked experiment framework. Treatment labels are saved as well.

2. When moving from stratum $i-1$ to stratum i, all factors applied in the higher strata are now treated as "whole-plot" factors. The corresponding model matrix

$\boldsymbol{Xw.m}$ containing terms inherited from the higher strata is expanded accordingly, as treatments applied at each unit in stratum $i-1$ are now applied to n_i units of the current stratum nested within it. A similar expansion procedure is carried out for the vector (or matrix, if there are two or more higher strata with factors applied) of the treatment labels, so that for each current unit we are able to see what treatment has been applied to it at each stratum.

Blocking with no factors applied may occur at any stratum, not only at the first one. In such cases the procedure remains the same: skipping to the next stratum with some treatment applied, expanding the design matrices and vectors with treatment labels corresponding to the higher strata.

3. Once there is a model matrix with the "whole-plot" terms is formed, the search procedure might be started for the current stratum. The candidate set of treatments is set for the factors applied in the current stratum; parameters of interest include not only the ones formed by these factors but also their interactions with the higher strata terms. As nesting within the previous strata is treated as fixed block effects, the criteria used are the ones given in (**??**) and (**??**). However, there are a few features worth noting:

   - For each design under consideration during the extensive search procedure, its model matrix $\boldsymbol{X}_1$ is now constructed by binding the "whole-plot" model matrix $\boldsymbol{Xw.m}$, the model comprising the terms formed from the factors applied at the current stratum $\boldsymbol{Xi.m}$, and the matrix formed of the interaction terms (if any are to be included) between the two. The same relates to the construction of the potential terms matrix $\boldsymbol{X}_2$: it needs to be recalculated every time a design point is swapped between the candidate set of the current stratum terms and the current design if it contains any interactions involving terms inherited from the previous strata. If not, it then only comprises terms from the current stratum factors.

   - Presence of the potential terms matrix in the criteria also implies that each stratum i will "have" its own number of potential terms q_i and, therefore, in the cases when the value of the variance scaling parameter τ^2 depends on it, at each stratum the criterion function will be evaluated with the respective values of τ_i^2 instead of some common one for all levels. In this work we consider common values of τ^2, however, it is a case-sensitive parameter, and it is to be discussed in each particular case.

   - As the numbers of primary and potential terms vary from stratum to stratum, so do the significance levels α_{LP} and α_{LoF} in the case of trace-based criterion (**??**):

$$\alpha_{LP} = 1 - (1 - \alpha_1)^{\frac{1}{p}},$$
$$\alpha_{LoF} = 1 - (1 - \alpha_2)^{\frac{1}{q}},$$

as the corrected confidence levels depend on the dimension of the confidence regions (as in (**??**)).

- We use the same values of weights in the criteria for all strata; however, the flexibility of the algorithm allows changing weights (and even criteria) between the strata.

4. If there are at least 3 strata with some factors applied, and when the current stratum number is 3 or more, an additional swapping procedure is performed (the same as that described by Trinca and Gilmour (2001)). By looking at the $i-2$ stratum units that have the same treatments applied to them, and interchange the $i-1$ stratum units within those, the performance of the design evaluated with respect to the performance at the current stratum i. The same swapping is performed for all the higher strata up to the first one.

5. It is all then repeated from step number 2, until current stratum i reaches the lowest stratum s.

References

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

Bailey, R. A. (2008) *Design of Comparative Experiments*, vol. 25. Cambridge University Press.

Box, G. E. P. and Draper, N. R. (1987) *Empirical Model-Building and Response Surfaces*. John Wiley & Sons.

Box, G. E. P. and Wilson, K. B. (1951) On the experimental attainment of optimum conditions (with discussion). *Journal of the Royal Statistical Society, Series B*, **13**, 1–45.

Cornell, J. A. (2011) *Experiments with Mixtures: Designs, Models, and the Analysis of Mixture Data*. John Wiley & Sons.

Cox, D. R. (1958) *Planning of Experiments*. John Wiley & Sons.

Draper, N. R. and Smith, H. (1998) *Applied Regression Analysis, 3rd edn*. John Wiley & Sons.

DuMouchel, W. and Jones, B. (1994) A simple Bayesian modification of D-optimal designs to reduce dependence on an assumed model. *Technometrics*, **36**, 37–47.

Fedorov, V. V. (1972) *Theory of Optimal Experiments*. Elsevier.

Gilmour, S. G. and Goos, P. (2009) Analysis of data from non-orthogonal multistratum designs in industrial experiments. *Journal of the Royal Statistical Society: Series C*, **58**, 467–484.

Gilmour, S. G., Goos, P. and Großmann, H. (2017a) Distribution-robust methods for analyzing data from split-plot and multi-stratum designs. In preparation.

Gilmour, S. G., Goos, P. and Großmann, H. (2017b) Pure error REML for analyzing data from split-plot and multi-stratum designs. In preparation.

Gilmour, S. G. and Trinca, L. A. (2000) Some practical advice on polynomial regression analysis from blocked response surface designs. *Communications in Statistics-Theory and Methods*, **29**, 2157–2180.

Gilmour, S. G. and Trinca, L. A. (2012) Optimum design of experiments for statistical inference (with discussion). *Journal of the Royal Statistical Society: Series C*, **61**, 345–401.

Goos, P., Kobilinsky, A., O'Brien, T. E. and Vandebroek, M. (2005) Model-robust and model-sensitive designs. *Computational Statistics & Data Analysis*, **49**, 201–216.

Goos, P., Langhans, I. and Vandebroek, M. (2006) Practical inference from industrial split-plot designs. *Journal of Quality Technology*, **38**, 162–179.

Harville, D. A. (2006) *Matrix Algebra From a Statistician's Perspective*. Springer New York.

Hinkelmann, K. and Kempthorne, O. (2005) *Design and Analysis of Experiments. Volume 2: Advanced Experimental Design*. John Wiley & Sons.

Kenward, M. G. and Roger, J. H. (1997) Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 983–997.

Koch, K. R. (2007) *Introduction to Bayesian Statistics*. Springer, 2nd edn.

Letsinger, J. D., Myers, R. H. and Lentner, M. (1996) Response surface methods for bi-randomization structures. *Journal of Quality Technology*, **28**, 381–397.

Lu, L., Anderson-Cook, C. M. and Robinson, T. J. (2011) Optimization of designed experiments based on multiple criteria utilizing a Pareto frontier. *Technometrics*, **53**, 353–365.

McCulloch, C. E. and Searle, S. R. (2001) *Generalized, Linear and Mixed Models*. John Wiley & Sons.

Mead, R., Gilmour, S. G. and Mead, A. (2012) *Statistical Principles for the Design of Experiments: Applications to Real Experiments*. Cambridge University Press.

Montepiedra, G. and Fedorov, V. V. (1997) Minimum bias designs with constraints. *Journal of Statistical Planning and Inference*, **63**, 97–111.

Myers, R. H., Montgomery, D. C. and Anderson-Cook, C. M. (2009) *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. John Wiley & Sons.

Rudin, W. (1987) *Real and Complex Analysis*. Tata McGraw-Hill Education.

Stallings, J. W. and Morgan, J. P. (2015) General weighted optimality of designed experiments. *Biometrika*, 1–11.

Trinca, L. A. and Gilmour, S. G. (2001) Multistratum response surface designs. *Technometrics*, **43**, 25–33.

Trinca, L. A. and Gilmour, S. G. (2015) Improved split-plot and multistratum designs. *Technometrics*, **57**, 145–154.

Yates, F. (1939) The recovery of inter-block information in variety trials arranged in three-dimensional lattices. *Annals of Eugenics*, **9**, 136–156.