Scientific Research Publishing

American Journal of Computational Mathematics, 2022, 12, 216-231
https://www.scirp.org/journal/ajcm
ISSN Online: 2161-1211
ISSN Print: 2161-1203

Variable Selection in Randomized Block Design Experiment

Sadiah Mohammed Aljeddani

Department of Mathematics, Umm Al Qura University, Makkah, Saudi Arabia
Email: smgadany@uqu.edu.sa

How to cite this paper: Aljeddani, S.M. (2022) Variable Selection in Randomized Block Design Experiment. *American Journal of Computational Mathematics*, **12**, 216-231.
https://doi.org/10.4236/ajcm.2022.122013

Received: April 24, 2022
Accepted: June 7, 2022
Published: June 10, 2022

Abstract

In the experimental field, researchers need very often to select the best subset model as well as reach the best model estimation simultaneously. Selecting the best subset of variables will improve the prediction accuracy as noninformative variables will be removed. Having a model with high prediction accuracy allows the researchers to use the model for future forecasting. In this paper, we investigate the differences between various variable selection methods. The aim is to compare the analysis of the frequentist methodology (the backward elimination), penalised shrinkage method (the Adaptive LASSO) and the Least Angle Regression (LARS) for selecting the active variables for data produced by the blocked design experiment. The result of the comparative study supports the utilization of the LARS method for statistical analysis of data from blocked experiments.

Keywords

Variable Selection, Shrinkage Methods, Linear Mixed Model, Blocked Designs

1. Introduction

Variable selection and statistical modelling are extremely useful in two aspects: prediction accuracy and clear interpretation. Prediction accuracy can be improved by removing the un-significant variables or shrinking them toward zero to have almost no effect on the response surface variable. This method reduces the variance of the predicted values and improves the overall prediction accuracy. Moreover, we often like to have a smaller and more meaningful model for good interpretation. Ensuring the selected model has the highest prediction accuracy will allow the model to be used for future predictions. Parallel to this, selecting the simplest model by avoiding uninformative variables, which often do not in-

fluence the response variable, is also crucial to enhancing scientific analysis. By removing the noninformative or nonactive variables, the predictive ability of models can be improved and parsimoniously describe the relationship between the informative, or active variables and the response variable. Variable selection issue refers to obtaining an adequate subset of variables for the model. The subject of variable selection in linear regression analysis is a remarkable subject. The experimenter initially may be uncertain about the most influential structure of the model. It might be unclear whether all the variables should be included in the model or if only some of them have significant effects on the response variable. Therefore, the variable selection procedure builds a regression model with an appropriate subset of variables.

As part of the former studies in this field, [1] and [2] proposed a shrinkage technique. They defined the regularisation methods by adding a penalty function to the residual sum of squares, and minimization or maximization of the penalised function with respect to the coefficients yields penalised likelihood estimators. Moreover, [3] discussed the penalised method for the linear mixed model. In their study, they dealt with supersaturated designs as they are very cost-effective with respect to the number of runs. This kind of design is very desirable in industrial experiments. They introduced a nonconvex penalised least squares approach. However, their study added the penalty function to the least square estimator while they deal with supersaturated experiments in which correlated data is expected. In [4] and [5], they overcome the problem of correlated data and selecting variables in a linear mixed model context by adding the penalty function to the generalised least square (GLS) estimator in which the two variance components will be estimated by the restricted maximum likelihood (REML) estimator. Furthermore, [6] developed the method of block deletion and block addition to fit high-dimensional data sets. They developed the threshold method. In their approach, they set a threshold value with the approximation error evaluated using the data. While the variable selection process is running, the error will be approximated and updated for the new model. The next step will use the updated threshold value and process the variable selection again. This updating prevents deleting useful variables. The aim of this paper is to compare the analysis of the frequentist methodology (the backward elimination), penalised shrinkage method (the Adaptive LASSO) and the Least Angle Regression (LARS) for selecting the active variables for data produced by the blocked design experiment. A simulation study using the design of the blocked experiment was also applied to support the comparative study.

The work by [4] differs from this work in which they did not apply the Adaptive LASSO to the pastry dough experiment. Also, the work by [5] has not studied the blocked design and the pastry dough experiment.

2. Linear Mixed Model and Analysis

The model for the block experiments includes two types of errors: block error

and residual error. Hence, linear mixed models (LMMs) are used to analyse responses from the blocked experiments.

Linear mixed-effects models (LMMs) introduce correlations between observations using random effects. This leads to the use of generalised least squares (GLS) estimation, combined with restricted maximum likelihood estimation (REML) of the variance components as will be discussed. This type of analysis is used by the most design of experiments textbooks that deal with blocked designs. In matrix notation, the model corresponding to a blocked design is written as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{Y}$ is $n \times 1$ vector of observations on the response of interest, $\boldsymbol{X}$ is the $n \times p$ model design matrix containing the polynomial expansions of the m factor levels at the n experimental runs, $\boldsymbol{\beta}$ is the $p \times 1$ vector of unknown fixed parameters, $\boldsymbol{Z}$ is an $n \times b$ random design matrix which represents the allocation of the runs to blocks, and whose (i, j)th element is one where the ith observation belongs to the jth blocks, and zero otherwise. If the runs of the experiment are grouped per block, then $\boldsymbol{Z}$ is of the form

$$\boldsymbol{Z} = \operatorname{diag}\left[\mathbf{1}_{k_1}, \mathbf{1}_{k_2}, \cdots, \mathbf{1}_{k_b}\right], \tag{2}$$

where $\mathbf{1}_k$ is a k vector of ones, and $k_1, k_2, \cdots, k_b$ are the blocks sizes. The random effects of the b blocks are contained within the $b \times 1$ vector $\boldsymbol{\gamma}$, and the random errors are contained within the $n \times 1$ vector $\boldsymbol{\epsilon}$. It is assumed that $\boldsymbol{\gamma}$ and $\boldsymbol{\epsilon}$ are independent and normally distributed, *i.e.* $\operatorname{Cov}(\boldsymbol{\gamma}, \boldsymbol{\epsilon}) = \mathbf{0}_{b \times n}$, where $\mathbf{0}_{b \times n}$ is the $b \times n$ matrix of zeros. Hence, $\boldsymbol{\gamma} \sim \mathrm{N}\left(\mathbf{0}_b, \sigma_\gamma^2 \boldsymbol{I}_b\right)$, and $\boldsymbol{\epsilon} \sim \mathrm{N}\left(\mathbf{0}_n, \sigma_\epsilon^2 \boldsymbol{I}_n\right)$, where $\mathbf{0}_b$ and $\mathbf{0}_n$ are the b and n column vectors of zeros respectively, and $\boldsymbol{I}_b$ and $\boldsymbol{I}_n$ are the b-dimensional and n-dimensional identity matrices respectively.

Under these assumptions, $\boldsymbol{Y}$ is a normally distributed random variable with mean $\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$, and the variance-covariance matrix of the response $\boldsymbol{Y}$ can be written as

$$\boldsymbol{V} = \operatorname{Var}(\boldsymbol{Y}) = \operatorname{Var}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}) \tag{3}$$

$$= \operatorname{Var}(\boldsymbol{Z}\boldsymbol{\gamma}) + \operatorname{Var}(\boldsymbol{\epsilon}) \tag{4}$$

$$= \boldsymbol{Z}\operatorname{Var}(\boldsymbol{\gamma})\boldsymbol{Z}' + \sigma_\epsilon^2 \boldsymbol{I}_n \tag{5}$$

$$= \sigma_\gamma^2 \boldsymbol{Z}\boldsymbol{Z}' + \sigma_\epsilon^2 \boldsymbol{I}_n. \tag{6}$$

$\boldsymbol{V}$ can be given as a block diagonal,

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V}_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{V}_2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \boldsymbol{V}_b \end{bmatrix},$$

where

$$\boldsymbol{V}_i = \sigma_\epsilon^2 \boldsymbol{I}_{k_i} + \sigma_\gamma^2 \mathbf{1}_{k_i} \mathbf{1}_{k_i}',$$

and

$$\boldsymbol{V}_i = \begin{bmatrix} \sigma_\varepsilon^2 + \sigma_\gamma^2 & \sigma_\gamma^2 & \cdots & \sigma_\gamma^2 \\ \sigma_\gamma^2 & \sigma_\varepsilon^2 + \sigma_\gamma^2 & \cdots & \sigma_\gamma^2 \\ \vdots & \ddots & \ddots & \vdots \\ \sigma_\gamma^2 & \cdots & \sigma_\gamma^2 & \sigma_\varepsilon^2 + \sigma_\gamma^2 \end{bmatrix}.$$

As a result, the variance-covariance matrix $\boldsymbol{V}_i$ of all observations within one block is compound symmetric: the main diagonal of the matrix contains the variances of the observations, while the off-diagonal elements are covariances. However, $\boldsymbol{V}_i$ can be rewritten as

$$V_i = \sigma_\epsilon^2 \left(\boldsymbol{I}_{k_i \times k_i} + \frac{\sigma_\gamma^2}{\sigma_\epsilon^2} \mathbf{1}_{k_i} \mathbf{1}'_{k_i} \right) \tag{7}$$

$$= \sigma_\epsilon^2 \left(\boldsymbol{I}_n + \eta \boldsymbol{Z}\boldsymbol{Z}' \right), \tag{8}$$

where $\eta = \sigma_\gamma^2 / \sigma_\epsilon^2$ is a measure for the extent to which observations within the same block are correlated. The larger this variance ratio, the stronger observations within the same block are correlated.

When the random error terms as well as the group effects are normally distributed, the maximum likelihood estimate of the unknown model parameter $\boldsymbol{\beta}$ in Equation (1) is the generalised least squares (GLS) estimate. Detecting the estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, requires to minimise

$$\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \right)' \boldsymbol{V}^{-1} \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \right) = \boldsymbol{y}'\boldsymbol{V}^{-1}\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{\beta} \tag{9}$$

with respect to $\boldsymbol{\beta}$, which is tantamount to detecting $\hat{\boldsymbol{\beta}}$, so that

$$\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X} \right) \hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}. \tag{10}$$

Therefore, the generalised least squares (GLS) estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X} \right)^{-1} \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y}, \tag{11}$$

and the variance-covariance matrix of the estimators is given by

$$\begin{aligned} \mathrm{Var}\left(\hat{\boldsymbol{\beta}} \right) &= \mathrm{Var}\left(\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X} \right)^{-1} \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y} \right) \right) \\ &= \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X} \right)^{-1} \boldsymbol{X}'\boldsymbol{V}^{-1} \mathrm{Var}\left(\boldsymbol{Y} \right) \left(\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X} \right)^{-1} \boldsymbol{X}'\boldsymbol{V}^{-1} \right)' \\ &= \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X} \right)^{-1} \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{V}^{-1}\boldsymbol{X} \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X} \right)^{-1} \\ &= \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X} \right)^{-1} \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X} \right) \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X} \right)^{-1} \\ &= \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X} \right)^{-1}. \end{aligned} \tag{12}$$

Often, the variances σ_γ^2 and σ_ϵ^2 are not known and therefore, Equation (11) and Equation (12) cannot be used directly. Instead, the estimates of the variance components, $\hat{\sigma}_\gamma^2$ and $\hat{\sigma}_\epsilon^2$, are substituted in the GLS estimator as in Equation (11), yielding

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X} \right)^{-1} \boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{Y}, \tag{13}$$

where

$$\boldsymbol{V} = \hat{\sigma}_{\epsilon}^{2}\boldsymbol{I}_{n} + \hat{\sigma}_{\gamma}^{2}\boldsymbol{Z}\boldsymbol{Z}'. \tag{14}$$

In that case, the variance-covariance matrix in Equation (12) can be approximated by

$$\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right) = \left(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}. \tag{15}$$

The generalised least square (GLS) estimator is unbiased, meaning that $\mathbb{E}\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}$, and is equal to the maximum likelihood estimator (MLE). The likelihood function defined as it is the joint probability density function for the observed data examined as a function of the parameters. Hence, the likelihood function for $\boldsymbol{Y}$ in Equation (1) is

$$L\left(\boldsymbol{\beta} \mid \boldsymbol{Y}\right) = \left(2\pi\right)^{-n/2}\left|\boldsymbol{V}\right|^{-1/2}\exp\left[-\frac{1}{2}\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)'\boldsymbol{V}^{-1}\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)\right], \tag{16}$$

where π is a constant which does not depend on $\boldsymbol{\beta}$. The maximum likelihood estimator (MLE) is the estimator that maximises the likelihood function, which is tantamount to detecting the $\hat{\boldsymbol{\beta}}$ as

$$\frac{\partial}{\partial\boldsymbol{\beta}}L\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{Y}\right) = 0, \tag{17}$$

which is equal to

$$\frac{\partial}{\partial\boldsymbol{\beta}}\ln L\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{Y}\right) = 0, \tag{18}$$

where $\ln L\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{Y}\right)$ is the log likelihood function. As Equation (9) is proportionate to log of Equation (16), the GLS estimator in Equation (11) is the result of Equation (17) and Equation (18).

Moreover, $\boldsymbol{V}$ can be estimated when observed data is obtained. In this work, we used the Restricted Maximum Likelihood (REML) estimator to estimate $\boldsymbol{V}$. According to [7], "REML requires the transformation of the response to remove the influence of the other model parameters followed by the maximisation of the likelihood for these transformed responses". The likelihood in REML includes knowledge about the variance components yet does not include knowledge about the fixed effects [8].

The restricted maximum likelihood (REML) used to estimate σ_{ϵ}^{2} and σ_{γ}^{2} is

$$l_{\mathrm{REML}}\left(\sigma_{\epsilon}^{2},\sigma_{\gamma}^{2};\boldsymbol{Y}\right) = -\frac{1}{2}\ln\left|\boldsymbol{V}\right| - \frac{1}{2}\ln\left|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right| - \frac{1}{2}\left(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\boldsymbol{V}^{-1}\left(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right), \tag{19}$$

where $\hat{\boldsymbol{\beta}}$ is defined in Equation (13). The restricted log-likelihood $l_{\mathrm{REML}}\left(\sigma_{\epsilon}^{2},\sigma_{\gamma}^{2};\boldsymbol{Y}\right)$ is minimised with respect to the variance components σ_{ϵ}^{2} and σ_{γ}^{2} to obtain an unbiased estimate for the variance components. In this work, REML is minimised by using the function "fmincon" in Matlab.

3. The Backward Elimination Method

Backward elimination starts with the full model, and then the least significant

variable, which corresponds to the highest p-value above a significance level α, will be dropped. The reduced models in each step are re-fitted by following the rule of significance level until all remaining variables are statistically significant meaning the corresponding variable has a p-value $\leq \alpha$.

We did the generalised least square estimator using the backward elimination to compare with other methods. The relative theory of the Backward elimination starts with the full model and eliminates one variable at a time based on the Wald (**Wa**) test statistic. The Wald test statistic is suitable to compare nested models when the variance-covariance matrix changes after each drop. We estimate the variance components σ_{ϵ}^2 and σ_{γ}^2 by REML using the full model, so we deal with the Wald test statistic as the variance components are known. We used the Wald test statistic at the 5% significance level, and we compare nested models through the process of backward elimination. Assuming that $\boldsymbol{y}$ is normally distributed and the variance components are known, we test the hypothesis:

$$H_0 : \beta_i = 0 \text{ vs } H_1 : \beta_i \neq 0$$

The Wald (**Wa**) test statistic will follow an F distribution with 1 and r degrees of freedom.

$$\mathbf{Wa} = \frac{\hat{\beta}_i^2}{\sigma_{ii}^2} \sim F_{(1,r)} \tag{20}$$

where σ_{ii}^2 is the (i,i) th element of $\text{Var}\left(\hat{\boldsymbol{\beta}}\right)$. The degrees of freedom in a statistical calculation represents how many values are involved in a calculation that has the freedom to vary. The degrees of freedom can be defined as they are equal to the number of independent observations minus the number of parameters. The degrees of freedom could be calculated to guarantee the statistical accuracy of tests statistics such as chi-square tests, t-tests and F-tests. Often, these tests are utilized to make a comparison between the observed data with the data that could be expected to be achieved according to a particular hypothesis.

4. Adaptive LASSO Shrinkage Method (ALASSO)

To overcome the drawbacks in the classical approaches of variable selection, [1] and [2] proposed regression modelling by regularisation technique. This technique prevents overfitting by restricting the model, typically to reduce its complexity. The regularisation methods are based on shrinkage penalties, where penalty functions are added to the residual sum of squares or subtracted from the log-likelihood, and minimisation or maximisation of penalised functions with respect to coefficients yields penalised likelihood estimators. The shrinkage penalty method can be explained as there is a penalty for any nonzero estimate of the model when we minimise the sum of the squared residuals. Thus, the penalty will shrink the size of the estimated coefficients toward zero. It places a constraint on the size of the regression coefficients [9] [10]. Shrinkage methods do not explicitly select variables, instead, they minimise the sum of the squared re-

siduals by applying a penalty on the size of the estimated coefficients. They have the advantage of selecting variables and estimating the coefficients simultaneously. The advantage of shrinkage methods is that their use often improves the prediction accuracy and helps with the selection of a more parsimonious model, though there is a trade-off between bias and the variance of the final model (see [11] [12]).

Based on the size of the estimated coefficients, the penalised estimates might be only shrunk in the size while in the case of small estimated coefficients, it is more likely the penalised estimates will be set to zero. Hence, the choice of the shrinkage parameter λ is sensitive and important. Unlike traditional subset selection, penalised regression is a continuous process as it shrinks the size of the coefficients and yields stable models with low prediction errors. However, as some shrinkage penalty functions shrink the size of the coefficients towards zero and not explicitly zero in the case of large size of estimated coefficients or the case of a very small amount of shrinkage parameter λ, the resulting models in such case suffer from complexity and overfitting.

For linear mixed models, the penalised generalised least squares estimates have been discussed by [3], it can be found by minimising

$$\boldsymbol{Q}_{\mathrm{PGLS}}(\boldsymbol{\beta}) = \frac{1}{2n}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})' \boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) + \sum_{j=1}^{d} p_{\lambda}\left(\left|\beta_j\right|\right), \tag{21}$$

with respect to $\boldsymbol{\beta}$, where $\hat{\boldsymbol{V}} = \hat{\sigma}_{\epsilon}^2 \boldsymbol{I}_n + \hat{\sigma}_{\gamma}^2 \boldsymbol{Z}\boldsymbol{Z}'$, and d is the number of the model coefficients $\boldsymbol{\beta}$. The expression $p_{\lambda}(.)$ is a penalty function and the shrinkage parameters λ is an unknown strictly positive thresholding parameter, which is often selected using information selection criteria after setting a grid for λ. In this work, we set a grid from 0 to 3. However, researchers can set any grid, this choice is suitable for the data that we used, and the effects size assumed in the simulation. It is assumed that $\mathbf{K} = 1 + p_t + v_c$, such that p_t is the number of the active fixed parameters in the fitted penalised least squares model [13]. The p_t can then be computed as $p_t = \mathrm{tr}\left[\boldsymbol{X}\left\{\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X} + \hat{\boldsymbol{W}}\right\}^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\right]$, where $\hat{\boldsymbol{W}}$ is a penalty matrix [3]. The v_c is the number of variance components that are used in fitting the penalised model. The following information selection criteria can be used in selecting the shrinkage parameters λ. Akaike Information Criterion (AIC) is given by,

$$\mathrm{AIC} = 2\mathbf{K} - 2\ln \hat{L}.$$

We illustrate the algorithim of the PGLS as follows:

1) Let $\boldsymbol{\beta}^{(0)}$ be the generalised least squares estimator $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ as in Equation (13), for the full model, a model fitted by all the variables in the experiment, and $\hat{\sigma}_{\epsilon}^2$, $\hat{\sigma}_{\gamma}^2$ be the REML estimates of the variance components for this model.

2) (a) Set a grid for λ with l values $\lambda_1, \lambda_2, \cdots, \lambda_l$ for each grid.

(b) For $i = 1, 2, \cdots, l$ of the grid, use λ_i to estimate the model parameters of the corresponding tuning.

(c) For $i = 1, 2, \cdots, l$ of the grid, choose λ_i that minimises $\mathrm{AIC}(\lambda_i)$ for λ.

(d) Return λ.

3) (a) Set $\boldsymbol{\beta}^{(0)}$ as the GLS estimator and λ given from the previous loop.

(b) We set the $\mathrm{SE}\left(\hat{\boldsymbol{\beta}}\right)$ as our proposed threshold to find out if the estimates are statistically significant or not, where $\mathrm{SE}\left(\hat{\beta}\right)=\sqrt{\mathrm{Var}\left(\hat{\beta}\right)}$ in Equation (15) is the standard error of $\hat{\boldsymbol{\beta}}$. Thus, we set values of $\hat{\boldsymbol{\beta}} \le \mathrm{SE}\left(\hat{\boldsymbol{\beta}}\right)$ equal to zero.

(c) All other (nonzero) coefficients are jointly updated using

$$\hat{\boldsymbol{\beta}}^{(1)}=\left\{\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}+\boldsymbol{W}^{(0)}\right\}^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{Y},$$

where $\hat{\boldsymbol{\beta}}^{(1)}$ is the vector of collecting all nonzero coefficients, and $\boldsymbol{W}^{(0)}$ is a penalty matrix for the initial values $\hat{\boldsymbol{\beta}}^{(0)}$ which it can be defined as

$$\boldsymbol{W}^{(0)}=\mathrm{diag}\left\{\frac{p'_{\lambda}\left(\left|\hat{\beta}_1^{(0)}\right|\right)}{\left|\hat{\beta}_1^{(0)}\right|},\cdots,\frac{p'_{\lambda}\left(\left|\hat{\beta}_{d^*}^{(0)}\right|\right)}{\left|\hat{\beta}_{d^*}^{(0)}\right|}\right\},$$

where the d^* is the total number of nonzero model coefficients.

(d) Any elements of $\hat{\boldsymbol{\beta}}^{(1)}$ that are $\hat{\boldsymbol{\beta}} \le \mathrm{SE}\left(\hat{\boldsymbol{\beta}}\right)$ are set to zero and the nonzero coefficients are jointly updated along with the matrix $\boldsymbol{W}^{(0)}$.

(e) Steps (c) to (d) are repeated until convergence takes place and no more factors can be removed.

4) Denote $\hat{\boldsymbol{\beta}}$ the final estimates of the nonzero model coefficients and $\hat{\boldsymbol{W}}$ the corresponding estimated $\boldsymbol{W}$ penalty matrix.

The covariance of the nonzero parameter estimates can then be obtained from the sandwich formula [3]:

$$\widehat{cov}\left(\hat{\boldsymbol{\beta}}\right)=\left(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}+\hat{\boldsymbol{W}}\right)^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\left(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}+\hat{\boldsymbol{W}}\right)^{-1} \tag{22}$$

The nonsignificant variables will be removed from the model, however, their indices will be saved and replaced with zero indicating that the variables have been removed.

[14] proposed the Adaptive LASSO (ALASSO) to obtain consistency in variable selection and prediction accuracy. ALASSO based on using a weighted (L_1), the LASSO penalty with weight determined by an initial estimator. Recall the PGLS in Equation (21), the penalty term by ALASSO is,

$$p_{\lambda}\left(\beta\right)=\lambda\sum_{j=1}^{d}\widehat{wg}_j\left|\beta_j\right|,$$

the vector of the weights $\widehat{wg}=\left(\widehat{wg}_1,\widehat{wg}_2,\cdots,\widehat{wg}_d\right)'$ are the adaptive data-driven weights, where $\widehat{wg}_j$, $j=1,\cdots,d$ can be constructed by $\widehat{wg}_j=\left(\hat{\beta}_j^{(0)}\right)^{-\psi}$, where ψ is a positive constant and $\hat{\boldsymbol{\beta}}^{(0)}$ is $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ [14].

To construct the adaptive weights $\widehat{wg}$, [14] suggested to pick a $\psi>0$. We take a fixed ψ such that $\psi>\frac{2\mu}{1-\mu}$, where $0\le\mu<1$. In our numerical studies, we let $\psi=\left[\frac{2\mu}{1-\mu}\right]+1$ as used by [15] to avoid the tuning of ψ. In their study, they concluded that for any $0\le\mu<1$, we can choose an appropriate ψ to

construct the adaptive weights and the oracle property holds as long as $\psi > \frac{2\mu}{1-\mu}$. These weights cause less shrinkage to more important predictors, which leads to consistent variable selection results. In this work, we set $\psi = 0.8$ as this choice was according our simulation study. We found that this choice provided the lowest Type I and II error rates among the choices that have been investigated during our research process.

5. Least Angle Regression Method (LARS)

Least Angle Regression (LARS) is a relative method proposed by [12], and can be viewed as a kind of "democratic" version of forward stepwise regression. The forward stepwise regression builds a model sequentially, adding one variable at a time. At each step, it identifies the best variable to include in a set which contains only active variables which considered to be statistically significant, and then update the least squares fit to include all the active variables.

The LARS algorithm has been described in [12]. Let $\boldsymbol{X}$ is the $n \times p$ design matrix for p factors. Let $\hat{\boldsymbol{\mu}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ be the LARS estimates of the response, and $\hat{\boldsymbol{\beta}}$ be the generalised least square estimate of the vector of coefficients. The LARS procedure works as follows:

1) Begin at $\hat{\mu}_0 = 0$ and set all coefficients to zero.

2) Find the variable, say x_1, which is most correlated with the response.

3) Fit the model using the generalised least square estimator in the direction of x_1 (or u_1, the unit vector along x_1) until another variable, say x_2, has as much correlation with the current residual as x_1 does.

4) At this point, the LARS estimate is updated to $\hat{\mu}_1 = \hat{\mu}_0 + \hat{\gamma}_1 u_1$, where $\hat{\gamma}_1$ is chosen such that the current residual $y - \hat{\mu}_1$ bisects the angle between x_1 and x_2.

5) Instead of continuing along x_1, LARS proceeds in the direction of u_2, the unit bisector of the two variables x_1 and x_2, until a third variable x_3 earns its way into the most correlated set.

6) Now the LARS estimate is updated to $\hat{\mu}_1 = \hat{\mu}_1 + \hat{\gamma}_2 u_2$, where $\hat{\gamma}_2$ is chosen such that the current residual $y - \hat{\mu}_2$ has equal angles with x_1, x_2, and x_3.

7) LARS then proceeds along u_3, the equiangular unit vector, *i.e.* along the least angle direction, until a fourth variable enters, etc. LARS builds up estimates in successive steps, in each step adding one variable to the model, so only p steps are required for the full set of solutions, where p is the number of variables.

The LARS algorithm in [12] provides the ordinary least square estimates. AljS modifed the LARS to deal with mixed effect models as follows: researcher needs to either weight the data as $\boldsymbol{X}^* = \sqrt{\boldsymbol{C}}\boldsymbol{X}$ and $\boldsymbol{Y}^* = \sqrt{\boldsymbol{C}}\boldsymbol{Y}$, where $\boldsymbol{C} = \boldsymbol{V}^{-1}$, and apply the lars function in the "lars" package in R. Othewise, researcher needs to modify the Gram matrix $\boldsymbol{G} = \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}$ to calculate the generalised least square estimator (GLS) in the lars function in Matlab.

6. Real Data Application

In this work, we demonstrate the differences of the variable selection methods which we studying on the nonorthogonally blocked response surface experiment. The real dataset for the pastry dough experiment has been used to apply the backward elimination, the Adaptive LASSO, and the LARS approaches. The estimated coefficients were obtained as well as the standard errors using all methods. The estimated variance components using REML in Equation (19), have been reported as well. We apply the methods in Sections 3, 4, and 5 using equations (20), (21), and the LARS algorithm to estimate the coefficients and the variance components. Also, we computed the GLS estimator using Equation (11). The standard errors are computed using the sandwich formula in Equation (22).

7. Analysis of the Pastry Dough Experiment

The design of the pastry dough experiment is nonorthoganally blocked response surface design. The design and the responses described in [16]. The factors invistigated in the experiment were the feed flow rate (x_1), the initial moisture content (x_2), the screw speed of mixing process for pastry dough (x_3). The purpose of the experimenter was to understand how the various properties of the dough depend on these three factors and how to develop an overall control scheme based on the experimental results. It was decieded that seven days of experimentation were affordable, so that 28 runs could be performed. We will apply our three variable selection methods on the response of measure the light transmission in bands of spectrum. The data of the experiment are given in **Table 1**. A full second order model in the three explonatory variables was used to explain the behavior of the response. The jth observation within the ith block can be expressed as

$$y_{ij} = \beta_0 + \sum_{i=1}^{3} \beta_i x_i + \sum_{i=1}^{3} \sum_{j=i+1}^{3} \beta_{ij} x_i x_j + \sum_{i=1}^{3} \beta_{ii} x_i^2 + \gamma_i + \varepsilon_{ij}$$

The variance compenents were estimated using the GLS-REML in the work by [17] as the $\hat{\sigma}_{\epsilon}^2 = 0.09695$ and $\hat{\sigma}_{\gamma}^2 = 0.9703$. The estimates for the ten parameters of the full quadratic model for the response are displayed in **Table 2**. The table contains the estimates obtained using the generlaised least square estimator, backward elimination, adaptive lasso, and least angle regression estimators. The REML has been used to estimate the variance components from the full quadratic model in this work. We found that both $\hat{\sigma}_{\epsilon}^2 = 0.09695$ and $\hat{\sigma}_{\gamma}^2 = 0.9702$, similar to the reults given by [17]. We note that the LARS method yield in a simillar model subset to the GLS in the original work by [17].

8. Simulation Study

To examine our methods, we need to run simulation studies in order to find out how the resulting model will be compared to the true model which used in the simulation. In the simulation, we set $\sigma_{\varepsilon}^2 + \sigma_{\gamma}^2 = 10$, and the variance components

Table 1. Data for the pastry dough mixing experiment.

obs.	Block	x_1	x_2	x_3	y
1	1	−1	−1	−1	12.92
2	1	−1	1	1	13.91
3	1	1	−1	1	11.66
4	1	1	1	−1	14.48
5	2	−1	−1	1	10.76
6	2	−1	1	−1	14.41
7	2	1	−1	−1	12.27
8	2	1	1	1	12.13
9	3	−1	1	−1	14.22
10	3	0	−1	0	12.35
11	3	1	0	0	13.50
12	3	0	0	1	12.54
13	4	1	−1	1	10.55
14	4	−1	0	0	13.33
15	4	0	1	0	13.84
16	4	0	0	−1	14.19
17	5	−1	−1	−1	11.46
18	5	1	1	1	11.32
19	5	0	0	0	11.93
20	5	0	0	0	11.63
21	6	−1	−1	1	12.20
22	6	1	1	−1	14.78
23	6	0	0	0	14.94
24	6	0	0	0	14.61
25	7	−1	1	1	12.17
26	7	1	−1	−1	11.28
27	7	0	0	0	11.85
28	7	0	0	0	11.64

Table 2. Estimated coefficients and standard errors (in parentheses) for the pastry dough experiment for y_1 by the Backward elimination, Adaptive Lasso (ALASSO) and the LARS. The last row is the estimated coefficients and p-values (in parentheses) from [17].

Method	β_{int}	β_{x_1}	β_{x_2}	β_{x_3}	$\beta_{x_1x_2}$	$\beta_{x_1x_3}$	$\beta_{x_2x_3}$	$\beta_{x_1^2}$	$\beta_{x_2^2}$	$\beta_{x_3^2}$
Backward	13.1263	−0.1894	0.8783	−0.7094	−0.1749	0	0	0	−0.5926	0
	(0.3821)	(0.0774)	(0.0775)	(0.0775)	(0.0925)	(−)	(−)	(−)	(0.1463)	(−)
ALASSO	13.1183	−0.1894	0.8783	−0.7092	−0.1874	0	0.1795	0	−0.4754	0
	(0.3633)	(0.0824)	(0.0746)	(0.0737)	(0.0875)	(−)	(0.0837)	(−)	(0.1598)	(−)
LARS	13.1249	−0.1894	0.8783	−0.7094	0	0	0	0	−0.5904	0
	(0.3634)	(0.0846)	(0.0846)	(0.0846)	(−)	(−)	(−)	(−)	(0.1713)	(−)
GLS	13.1960	−0.1894	0.8783	−0.7094	0	0	0	0	−0.4303	0
	(0.3910)	(0.0734)	(0.0734)	(0.0734)	(−)	(−)	(−)	(−)	(0.1877)	(−)

ratio to two different levels, $\eta = 1$ and 10. Similar values for η have been used for the analysis of data from many blocked and split-plot experiments (see, for instance, [18] and [16]). We generate 1000 datasets using the design structure from the motivating experiment in Table 1 for the pastry dough design. Given the assumed true model, we compare the performance of the backward elimination, ALASSO, and LARS. We focus on the properties of the estimated models by investigating the following properties:

1) Consistency in variable selection (frequency in selecting the active/nonactive variable).

2) Prediction performance.

For point 1, at 5% significant level, we report Type I error rate (an effect that is truly not significant but the corresponding procedure estimate indicates that it is significant). We also report Type II error rate (an effect that is truly present but the corresponding procedure estimate indicates that it is not significant).

For point 2, following [9] and [19], prediction accuracy is measured by computing the mean-squared error for each penalised estimate $\hat{\boldsymbol{\beta}}_\lambda$ as,

$$\mathbf{ME}\left(\hat{\boldsymbol{\beta}}_\lambda\right) = \left(\boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{X}\boldsymbol{\beta}\right)' \left(\boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{X}\boldsymbol{\beta}\right).$$

The relative model error (**RME**) is the ratio of the model error of the penalised estimates to the model error for the GLS estimates of the fixed effects,

$$\mathbf{RME} = \frac{\mathbf{ME}\left(\hat{\boldsymbol{\beta}}_\lambda\right)}{\mathbf{ME}\left(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}\right)},$$

where $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ in Equation (13) is the generalised least squares estimator of $\boldsymbol{\beta}$. The median of the relative model error (**MRME**) over 1000 simulated data sets were reported. **MRME** values greater than one indicate that the methods estimates perform worse than the GLS estimates, values near to one indicate that the the methods estimates performs in a similar way to the GLS estimates, values less than one indicate that the methods estimates performs better than the GLS estimates.

9. Simulation Study Using the Design of the Pastry Dough Experiment

A simulation study was performed to examine the performance of the backward, ALASSO and LARS estimates. Using the design of the pastry dough experiment from Table 1, the response variable was generated given the true model

$$\mathbb{E}\left(\boldsymbol{Y}\right) = 4x_1 + 2x_2 - 4x_1x_2 + 4x_1^2 + 2x_2^2.$$

In this experiment, five active variables x_1, x_2, x_1x_2, x_1^2 and x_2^2 and four nonactive variables x_3, x_1x_3, x_2x_3 and x_3^2 were assumed. We assumed this model as we would like to check a model with variety of factor types. Type I and II error rates of the design for the pastry dough experiments obtained using the backward, ALASSO, and LARS methods are given Table 3 and Table 4. Firstly, with

Table 3. Type I error rate for the pastry dough design.

True nonactive variable	η	x_3	x_1x_3	x_2x_3	x_3^2
Method		0	0	0	0
Backward	1	0.050	0.069	0.091	0.330
	10	0.045	0.050	0.089	0.131
ALASSO	1	0.050	0.066	0.077	0.265
	10	0.049	0.051	0.070	0.202
LARS	1	0.049	0.051	0.071	0.175
	10	0.040	0.046	0.050	0.096

Table 4. Type II error rate for the pastry dough design.

True active variable	η	x_1	x_2	x_1x_2	x_1^2	x_2^2
Method		4	2	−4	4	2
Backward	1	0.012	0.333	0.020	0.323	0.807
	10	0.011	0.301	0	0.005	0.774
ALASSO	1	0	0.220	0	0.166	0.884
	10	0	0.135	0	0.004	0.691
LARS	1	0	0.136	0	0.077	0.489
	10	0	0.098	0	0	0.409

regard to Type I error rate, for the main effects lie between 0.040 to 0.050 at both levels of η which is acceptable. The interaction effects are noticebly larger and run in the range of 0.046 to 0.091. The LARS method recorded the least Type I error rates at $\eta = 1$ and 10 for all effects. The quadratic effects yield in large Type I error rate in both levels $\eta = 1$ and 10 as this factor was hard to estimted as nonactive by all methods. However, the LARS yield in the lowest Type I error rate at $\eta = 1$ by rate of 0.175 and $\eta = 10$ by rate of 0.096. The quadratic effects are nearly not orthogonal to the blocks as the main effects and the two-factor interaction effects as made it difficult to estimate.

Secondly, with respect to the Type II error rate, the larger size of main and interaction effects was correctly estimated by almost no or few errors by all methods. In contrast, smaller size main and quadratic effects were hard to detect. However, the LARS method yields smaller Type II errors than the other methods. We notice that by the increase of the η value, the type error rates are reduced. This can be explained as the fact that the large value of η correspond to small values of σ_ϵ^2 as $\eta = \frac{\sigma_\gamma^2}{\sigma_\epsilon^2}$. This yields in smaller standard errors for each factor effect.

We found that the block design with correlated data affects the performance of the methods as well as the trade-off between Type I and Type II error rates. The analysis from the simulation study which used the nonorthogonal design

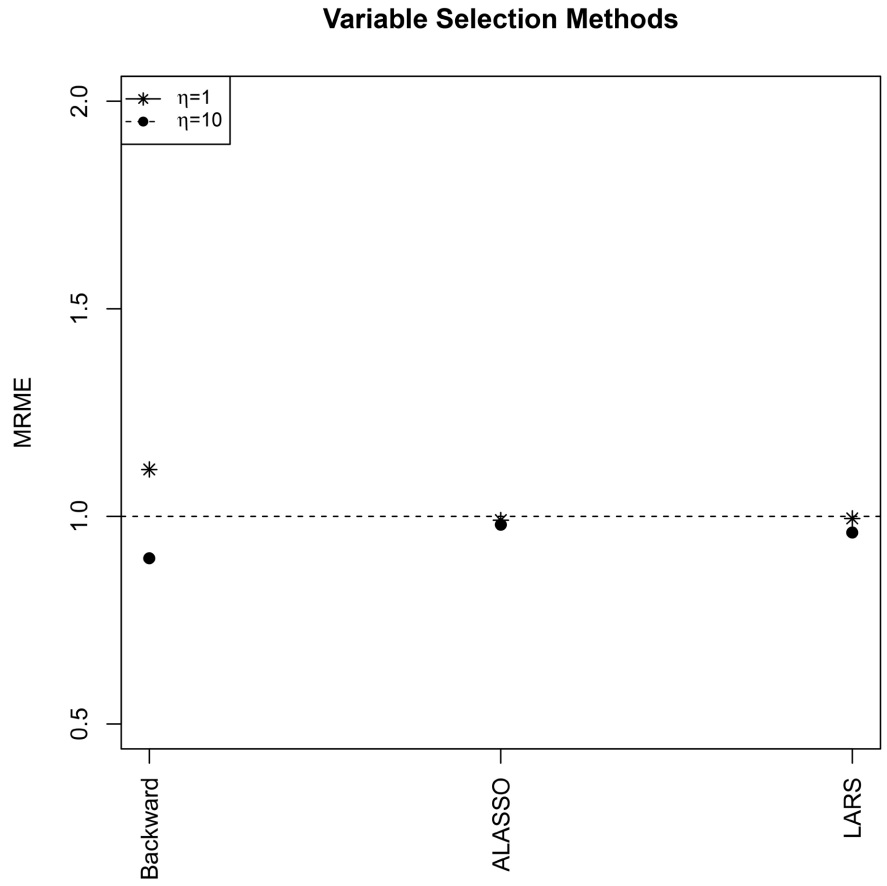

Figure 1. Median relative model error (MRME) for the pastry dough design.

from the pastry dough experiment showed that the LARS followed by ALASSO could control the Type II error rate at bot $\eta = 1$ and 10 better than the backward elimination. Furthermore, the Type I error rate was hard to control by backward and ALASSO at $\eta = 1$ and almost at $\eta = 10$ for all nonactive variables, especially for the quadratic effect x_3^2. The LARS at $\eta = 10$ performs better than the other methods with respect to controlling the Type I error rate.

Finally, with regard to the median relative model error (MRME), from **Figure 1**, we note that all methods have lower MRME than the GLS estimator method at $\eta = 10$. However, the LARS method has the lowest MRME. This indicates to the LARS has better performance than the GLS estimates. We conclude that the LARS method has the best performance among all methods used in the simulation.

10. Conclusion

This paper provided an analysis of data from blocked experiments using a motivating example from the industrial environment. Specifically, we recommend the use of the modified LARS method for variable selection in the blocked experiments. In our results, we observed that the LARS can identify the active variables (linear, two-factor interaction and quadratic), much better than the tradi-

tional used GLS method combined with backward elimination and the Adaptive LASSO shrinkage method. However, as expected this comes with the expense of slightly higher Type I error rates. We also observed a better prediction performance for the models chosen by the LARS compared to the models chosen by the other two methods.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Hoerl, A.E. and Kennard, R.W. (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**, 55-67. https://doi.org/10.1080/00401706.1970.10488634

[2] Breiman, L., *et al.* (1996) Heuristics of Instability and Stabilization in Model Selection. *The Annals of Statistics*, **24**, 2350-2383. https://doi.org/10.1214/aos/1032181158

[3] Li, R. and Lin, D.K.J. (2003) Analysis Methods for Supersaturated Design: Some Comparisons. *Journal of Data Science*, **1**, 249-260. https://doi.org/10.6339/JDS.2003.01(3).134

[4] Mylona, K. and Goos, P. (2011) Penalized Generalized Least Squares for Model Selection under Restricted Randomization. Manuscript Submitted for Isaac Newton Institute for Mathematical Sciences; NI 11032.

[5] Aljeddani, S. (2019) Statistical Analysis of Data from Experiments Subject to Restricted Randomisatio. PhD Thesis, University of Southampton, Southampton.

[6] Ozawa, S., Nagatani, T. and Abe, S. (2010) Fast Variable Selection by Block Addition and Block Deletion. *Journal of Intelligent Learning Systems and Applications*, **2**, 200-211. https://doi.org/10.4236/jilsa.2010.24023

[7] Matthews, E.S. (2015) *Design of Factorial Experiments in Blocks and Stages*. PhD Thesis, University of Southampton, Southampton.

[8] Corbeil, R.R. and Searle, S.R. (1976) Restricted Maximum Likelihood (Reml) Estimation of Variance Components in the Mixed Model. *Technometrics*, **18**, 31-38. https://doi.org/10.2307/1267913

[9] Fan, J.Q. and Li, R.Z. (2001) Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American statistical Association*, **96**, 1348-1360. https://doi.org/10.1198/016214501753382273

[10] Hastie, T., Tibshirani, R. and Friedman, J. (2009) Unsupervised Learning. In *The Elements of Statistical Learning*, Springer, New York, 485-585. https://doi.org/10.1007/978-0-387-84858-7_14

[11] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, *Series B* (*Methodological*), **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[12] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., *et al.* (2004) Least Angle Regression. *The Annals of Statistics*, **32**, 407-499. https://doi.org/10.1214/009053604000000067

[13] Wahba, G. (1980) Spline Bases, Regularization, and Generalized Cross Validation for Solving Approximation Problems with Large Quantities of Noisy Data. *Proceedings*

of the International Conference on Approximation Theory in Honor or George Lorenz, Austin, 8-11 January 1980.

[14] Zou, H. (2006) The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, **101**, 1418-1429. https://doi.org/10.1198/016214506000000735

[15] Zou, H. and Zhang, H.H. (2009) On the Adaptive Elastic-Net with a Diverging Number of Parameters. *Annals of Statistics*, **37**, 1733. https://doi.org/10.1214/08-AOS625

[16] 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. https://doi.org/10.1080/03610920008832601

[17] Goos, P. (2002) The Optimal Design of Blocked and Split-Plot Experiments. Springer Science & Business Media, New York. https://doi.org/10.1007/978-1-4613-0051-9

[18] Letsinger, J.D., Myers, R.H. and Lentner, M. (1996) Response Surface Methods for Birandomization Structures. *Journal of Quality Technology*, **28**, 381-397. https://doi.org/10.1080/00224065.1996.11979697

[19] Ibrahim, J.G., Zhu, H.T., Garcia, R.I. and Guo, R.X. (2011) Fixed and Random Effects Selection in Mixed Effects Models. *Biometrics*, **67**, 495-503. https://doi.org/10.1111/j.1541-0420.2010.01463.x