Penalized Generalized Least Squares for Model Selection under Restricted Randomization

Kalliopi Mylona[1,2]
Peter Goos[2,3]

[1]Southampton Statistical Sciences Research Institute,
University of Southampton, Southampton SO17 1BJ, United Kingdom.
[2]Faculty of Applied Economics, Universiteit Antwerpen,
City Campus - Prinsstraat 13, 2000 Antwerpen, Belgium.
[3]Erasmus School of Economics, Erasmus Universiteit Rotterdam,
Postbus 1738, 3000 DR Rotterdam, Netherlands.

mylona.kalliopi@ua.ac.be
peter.goos@ua.ac.be

Summary

For model selection purposes in experimental contexts, researchers often use stepwise regression or subset selection. With currently available software, this has to be done manually and often involves numerous model estimations in situations involving restricted randomization, such as block experiments and split-plot experiments. Moreover, these selection procedures ignore the stochastic errors inherited in the variable selection stage. This leads to incorrect standard errors. In this paper, we investigate the usefulness of penalized least squares estimation, which performs model selection and model estimation simultaneously. Therefore, the method results in correct standard errors. A key property of the penalized least squares estimation approach is that it possesses the so-called oracle property, which means that it works as well as if the correct sub-model were known. We study the performance of the approach using various practical examples, and investigate its properties in a simulation study.

Keywords: blocked experiments, hard thresholding penalty, L_1-penalty function, least absolute shrinkage and selection operator (LASSO), smoothly clipped absolute deviation (SCAD), split-plot design, variable selection.

1 Introduction

Response surface models often involve a relatively large number of regressors. Usually, not all the regressors have a significant impact on the responses of interest. One of the challenges for the researcher is therefore to select the best possible model. This is often done using stepwise regression (stepwise addition or stepwise deletion, or a combination of the two) or subset selection. For data from completely randomized experiments, such approaches are available in many commercial packages, and, as a result, they are frequently used. The statistical properties of these model selection procedures are, however, not well understood (for instance, because a sequence of tests is performed on the same data, so that there is an issue of multiple testing, and because p values are conditional on what variables have been included in or excluded from the model). Also, there is no obvious way to determine appropriate standard errors for the purpose of statistical inference and model prediction. This is due to the fact that these model selection procedures ignore the stochastic nature of the various steps of the variable selection.

Another drawback is that model selection using stepwise regression can be computationally cumbersome if there is a large number of regressors. This is because stepwise variable selection requires the estimation of a large number of models. On top of that, many experiments involve various responses, which increases the computational burden even further.

In industrial experimentation, there is an additional complication: many experiments involve one or more restrictions on the randomization. This leads to correlated observations and necessitates the use of generalized least squares (GLS) estimation, combined with restricted maximum likelihood estimation (REML) of the variance components. In that case, one can also carry out stepwise regression, but this has to be done manually as commercial packages do not offer this option in an automated form.

Consequently, there is a need for a method that simultaneously selects the regressors to be included in the model and estimates the model coefficients. This will not only facilitate the work to be done by the researcher (especially in the presence of multiple responses and in the case of a restricted randomization), but it will also make it possible to obtain reliable standard errors for the estimated model parameters and to circumvent the multiple testing issue.

We propose the use of penalized least squares for simultaneous model selection and estimation, which is a special case of penalized likelihood. Penalized least squares and penalized likelihood have already been used in a variety of contexts, including hazards and frailty models, high-dimensional knowledge discovery, analysis of supersaturated designs and large-scale medical studies (see Fan and Li (2001; 2002; 2006), Li and Lin (2002), Karagrigoriou *et al.* (2010), and Androulakis *et al.* (2010)). The main idea of penalized least squares is that there is a penalty for any non-zero estimate of the model coefficients when minimizing the sum of the squared residuals. As a result, the penalized least squares estimation approach has an incentive to estimate certain model parameters to zero. So, the penalized least squares approach has an incentive not to make the model unnecessarily complicated by including a large number of regressors. We show how the penalized least squares estimation approach can be adapted to cope with correlated observations from experiments with random block effects and from split-plot experiments, and we name the resulting approach penalized generalized least squares.

In the next section, we discuss the model, for which our approach will be used, for data from blocked and split-plot experiments. In Section 3, we provide a description of the traditional penalized least squares approach for uncorrelated observations, and present the most commonly used penalty functions. We also introduce the penalized generalized least squares approach. In Section 4, we demonstrate the usefulness of the penalized generalized least squares approach using four real-life data sets from the literature. In Section 5, we describe the results of an extensive simulation study which we carried out to investigate the properties of the penalized generalized least squares approach in more detail. Finally, we end the paper with a discussion.

2 Model

In this article, we study the linear model

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\nu},$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of responses, $\mathbf{X}$ is an $n \times p$ model matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector containing an intercept and $d = p - 1$ factor effects, and $\boldsymbol{\nu}$ is the vector of random errors. We denote the intercept by β_0 and the factor effects by $\beta_1, \ldots, \beta_d$. We focus on experiments involving randomly selected blocks and on split-plot experiments, in which the observations are obtained in groups. We denote the number of groups by b and the number of observations in each group by k. In blocked experiments, the different groups of observations are typically called blocks, whereas in split-plot experiments they are referred to as whole plots.

A key difference between blocked and split-plot experiments is that there are two sorts of factors in split-plot experiments. Some factors are held constant for all the observations within a group or whole plot, whereas others are reset independently for each individual observation. The former factors are called whole-plot factors, whereas the latter are referred to as sub-plot factors. The effects of the whole-plot factors are named whole-plot effects, while the effects of the sub-plot factors are called sub-plot effects. Interaction effects involving the two types of factors are called whole-plot-by-sub-plot interaction effects. Often, the levels of the whole-plot factors are, in some sense, hard to change, while the levels of the sub-plot factors are easy to change. When discussing split-plot designs in this article, we will denote the ith whole-plot factor by w_i and the ith sub-plot factor by s_i. In blocked experiments, all the factors are reset independently for each observation, so that there is only one type of experimental factor. Therefore, we denote the ith factor in a blocked experiment simply by x_i. Conceptually, the factors in a blocked experiment are similar to the sub-plot factors in a split-plot experiment.

For data from blocked and split-plot experiments, it is natural to assume that

$$\boldsymbol{\nu} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

where $\mathbf{Z}$ is an $n \times b$ matrix of zeros and ones whose (i, j)th element is one if the ith observation was obtained in group j, $\boldsymbol{\gamma}$ is a $b \times 1$ vector of random effects describing the group-to-group variation in the responses, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector containing the random errors for each of the n measured responses. The elements of $\boldsymbol{\gamma}$ and $\boldsymbol{\varepsilon}$ are assumed to be mutually independently normally distributed with zero mean and variances σ^2_γ and σ^2_ε respectively. The implied variance-covariance matrix for the response vector $\boldsymbol{Y}$ then is

$$\mathbf{V} = \sigma^2_\varepsilon \mathbf{I}_n + \sigma^2_\gamma \mathbf{Z}\mathbf{Z}' = \sigma^2_\varepsilon(\mathbf{I}_n + \eta\mathbf{Z}\mathbf{Z}'),$$

where $\eta = \sigma^2_\gamma/\sigma^2_\varepsilon$. The larger this variance ratio, the stronger observations within the same group are correlated.

The best linear unbiased estimator for the parameter vector $\boldsymbol{\beta}$ is the generalized least squares estimator

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

Using this expression requires the estimation of the variance components σ^2_γ and σ^2_ε. Gilmour and Trinca (2000) and Letsinger *et al.* (1996) recommend REML estimation of these variance components for data from blocked experiments and from split-plot experiments, respectively. Strong arguments in favor of REML estimates are that they have minimum variance in case the design is balanced and that, unlike other types of estimates, they are calculable for any type of blocked and split-plot designs, such as blocked and split-plot response surface designs (see, for instance, Searle *et al.* (1992)).

3 Penalized Least Squares

In this section, we introduce the penalized generalized least squares approach. First, we discuss the approach as it was originally proposed for models with uncorrelated observations. Next, we introduce the penalized generalized least squares approach to cope with correlated observations.

3.1 Penalty functions

Penalized least squares estimation differs from classical least squares estimation because the factor-effect estimates are obtained by minimizing an objective function that involves a penalty function on top of the sum of squared residuals. The idea of the penalty function is that it prevents overfitting: a penalty is incurred whenever a factor effect has a nonzero estimate, i.e. whenever a term is included in the model. Therefore, factor effects only get positive estimates if the resulting penalty is compensated by a substantial decrease in the sum of squared residuals. As a result, the estimate of any unimportant factor effect is zero. The penalized least squares estimates of the factor effects therefore automatically select an appropriate model.

For a linear model with uncorrelated errors, the penalized least squares estimates are obtained by minimizing

$$Q_{POLS}(\boldsymbol{\beta}) = \frac{1}{2}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}) + n\sum_{j=0}^{d} p_\lambda(|\beta_j|) \tag{2}$$

with respect to $\boldsymbol{\beta}$, instead of the sum of the squared residuals,

$$Q_{OLS}(\boldsymbol{\beta}) = (\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}), \tag{3}$$

which is minimized in ordinary least squares estimation. In expression (2), $p_\lambda(.)$ is a penalty function and λ is an unknown strictly positive thresholding parameter, which is often chosen using generalized cross-validation (Craven and Wahba (1979)).

In the literature, there are three commonly used penalty functions. The first of these is the L_1 penalty,

$$p_\lambda(|\beta|) = \lambda|\beta|,$$

which results in the least absolute shrinkage and selection operator (LASSO; see Tibshirani (1996)). The second is the hard thresholding penalty function,

$$p_\lambda(|\beta|) = \lambda^2 - (|\beta| - \lambda)^2 I(|\beta| < \lambda),$$

where $I(|\beta| < \lambda)$ is an indicator function that takes the value one if $|\beta| < \lambda$ and zero otherwise (Antoniadis (1997)). The final penalty function is the smoothly clipped absolute deviation (SCAD) penalty, proposed by Fan (1997),

$$p_\lambda(\beta) = \begin{cases} \lambda|\beta|, & \text{if } 0 \leq |\beta| < \lambda, \\ \frac{(\alpha^2-1)\lambda^2 - (|\beta|-\alpha\lambda)^2}{2(\alpha-1)}, & \text{if } \lambda \leq |\beta| < \alpha\lambda, \\ \frac{(\alpha+1)\lambda^2}{2}, & \text{if } |\beta| \geq \alpha\lambda. \end{cases}$$

For the tuning parameter α, Fan (1997) and Fan and Li (2001) suggest using a value of 3.7, because this value gave a satisfactory performance in a variety of variable selection problems.

Fan and Li (2001) show that the hard threshold penalty function and the SCAD penalty function have desirable theoretical properties, which make them ideal for variable selection, and that this is not the case for the L_1 penalty function. More specifically, the hard threshold and SCAD penalty functions possess the so-called oracle property, asymptotically. This means that with these penalty functions the penalized least squares estimator works as well as if the correct submodel were known in advance. In technical terms, this means that, when the true parameters have some zero components, they are estimated to zero with probability tending to one, and, when the true parameters are not zero, they are estimated as well as when the correct submodel is known. Furthermore, Fan and Li (2001) point out that the penalized least squares estimator improves the accuracy for estimating not only the zero parameters, but also for estimating the nonzero parameters, and that it outperforms the maximum likelihood estimator or GLS estimator. A simulation study indicated that the L_1 penalty function has good performance when the noise to signal ratio is large, but the bias created by this approach is noticeably large. The penalized least squares approach with the SCAD penalty function gives the best performance in selecting significant variables without creating excessive biases. The simulation study also demonstrated that the performance of SCAD is as good as that of the oracle estimator when the sample size n increases. In their conclusion, Fan and Li (2001) indicated that the penalized least squares approach can be applied to other statistical contexts without any extra difficulties. In Section 3.2, we show how the penalized least squares approach can be adapted to deal with correlated observations.

Fan and Li (2001) propose an iterative ridge regression estimation approach to obtain the penalized least squares estimates. The procedure starts from an initial value $\boldsymbol{\beta}^{(0)}$, which can, for instance, be the parameter estimates for the full model or the estimates from the model obtained after stepwise variable selection, as suggested by Li and Lin (2002). For each element $\beta_j^{(0)}$ of $\boldsymbol{\beta}^{(0)}$ which is very close to zero, the corresponding estimate is set to zero. All other (nonzero) coefficients are jointly updated using

$$\boldsymbol{\beta}^{(1)} = \{\mathbf{X}'\mathbf{X} + \mathbf{W}^{(0)}\}^{-1}\mathbf{X}'\boldsymbol{Y}, \tag{4}$$

where $\boldsymbol{\beta}^{(1)}$ is the vector collecting all nonzero coefficients and, if we denote by d^* the number of nonzero model coefficients,

$$\mathbf{W}^{(0)} = n \begin{bmatrix} \frac{p'_\lambda(|\beta_1^{(0)}|)}{|\beta_1^{(0)}|} & 0 & \dots & 0 \\ 0 & \frac{p'_\lambda(|\beta_2^{(0)}|)}{|\beta_2^{(0)}|} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{p'_\lambda(|\beta_{d^*}^{(0)}|)}{|\beta_{d^*}^{(0)}|} \end{bmatrix}.$$

Next, any elements of $\boldsymbol{\beta}^{(1)}$ that are close to zero are set to zero and the non-zero coefficients are jointly updated along with the matrix $\mathbf{W}^{(0)}$. These steps are repeated until convergence takes place. Because the parameter estimation and the variable selection are done simultaneously, the standard errors for the nonzero parameter estimates can be obtained directly from the sandwich formula

$$\widehat{cov}(\widehat{\boldsymbol{\beta}}) = \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X} + \widehat{\mathbf{W}})^{-1}\mathbf{X}'\mathbf{X}\{\mathbf{X}'\mathbf{X} + \widehat{\mathbf{W}}\}^{-1},$$

where $\widehat{\boldsymbol{\beta}}$ contains the final estimates of the nonzero model coefficients and $\widehat{\mathbf{W}}$ is the corresponding estimated $\mathbf{W}$ matrix.

The implementation of these methods requires the determination of the thresholding parameter λ. This can be done using generalized cross-validation. This method departs from the fact that

$$\hat{\boldsymbol{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}\{\mathbf{X}'\mathbf{X} + \widehat{\mathbf{W}}\}^{-\mathbf{1}}\mathbf{X}'\boldsymbol{Y},$$

which means that

$$\mathbf{X}\{\mathbf{X}'\mathbf{X} + \widehat{\mathbf{W}}\}^{-\mathbf{1}}\mathbf{X}'$$

plays the role of the hat matrix. The quantity

$$e = tr[\mathbf{X}\{\mathbf{X}'\mathbf{X} + \widehat{\mathbf{W}}\}^{-\mathbf{1}}\mathbf{X}']$$

therefore measures the number of effective parameters in the fitted penalized least squares model. The generalized cross-validation statistic is then

$$\mathrm{GCV}(\lambda) = \frac{1}{n}\frac{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2}{(1 - e/n)^2}.$$

Note that this statistic depends on λ through $\hat{\boldsymbol{Y}}$ and through the number of effective parameters, e. The thresholding parameter is chosen so that this statistic is minimized.

3.2 Correlated observations

In this section, we present two different modifications of the penalized least squares approach. The first modification treats all the factor effects alike, and works well for blocked experiments. The second modification is suitable for split-plot experiments, which have two types of factors whose effects are estimated with different precisions. For this reason, we used different penalties for whole-plot and sub-plot factor effects.

3.2.1 Blocked experiments

In the case of blocked experiments, the penalized least squares estimates can be found by minimizing

$$Q_{PGLS}(\boldsymbol{\beta}) = \frac{1}{2}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-\mathbf{1}}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}) + n\sum_{j=0}^{d} p_\lambda(|\beta_j|) \qquad (5)$$

with respect to $\boldsymbol{\beta}$, instead of the weighted sum of squared residuals

$$Q_{GLS}(\boldsymbol{\beta}) = (\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}),$$

which would produce the generalized least squares estimates. In this paper, we implemented the L_1 penalty function (LASSO, Tibshirani (1996)), the hard thresholding penalty function (HARD, Antoniadis (1997)) and the smoothly clipped absolute deviation (SCAD) penalty function with $\alpha = 3.7$ (Fan and Li (2001)).

To obtain the penalized generalized least squares estimates, we modified the estimation procedure from the previous section to deal with correlated observations. We updated the (nonzero) coefficients jointly using

$$\boldsymbol{\beta}^{(1)} = \{\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} + \mathbf{W}^{(0)}\}^{-1}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{Y}. \tag{6}$$

As in the uncorrelated case, these steps are repeated until convergence takes place. The standard errors for the nonzero parameter estimates can then be obtained from the sandwich formula

$$\widehat{cov}(\widehat{\boldsymbol{\beta}}) = \{\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} + \widehat{\mathbf{W}}\}^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\{\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} + \widehat{\mathbf{W}}\}^{-1},$$

where $\widehat{\boldsymbol{\beta}}$ again contains the final estimates of the nonzero model coefficients and $\widehat{\mathbf{W}}$ is the corresponding estimated $\mathbf{W}$ matrix.

We also used a generalized cross-validation to estimate the thresholding parameter λ. We computed the generalized cross-validation statistic as

$$\mathrm{GCV}(\lambda) = \frac{1}{n}\frac{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2}{(1 - e/n)^2},$$

where

$$\hat{\boldsymbol{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} + \widehat{\mathbf{W}})^{-1}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{Y}$$

and

$$e = tr[\mathbf{X}\{\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} + \widehat{\mathbf{W}}\}^{-1}\mathbf{X}'\mathbf{V}^{-1}],$$

and used the value of λ that minimizes $\mathrm{GCV}(\lambda)$. At each step of the estimation procedure, the variance components σ^2_γ and σ^2_ε are re-estimated using REML.

3.2.2 Split-plot experiments

Equation (5) works well for blocked experiments but not for split-plot experiments. This is because the method is not capable of keeping the type I error rates for the main effects, the quadratic effects and the interaction effects of the whole-plot factors under control. This can be remediated by using larger penalties for whole-plot factor effects. Therefore, for analyzing data from split-plot experiments, we use separate penalties for the whole-plot and sub-plot effects. The penalized least squares estimates are found by minimizing

$$Q_{PGLS}(\boldsymbol{\beta}) = \frac{1}{2}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}) + \sum_{j=0}^{d_w} p_{\lambda_w}(|\beta_j|) + n \sum_{j=d_w+1}^{d} p_{\lambda_s}(|\beta_j|) \quad (7)$$

with respect to $\boldsymbol{\beta}$. In this expression, d_w denotes the number of whole-plot effects, $\beta_1, \dots, \beta_{d_w}$ represent these whole-plot effects and $\beta_{d_w+1}, \dots, \beta_d$ are the sub-plot effects. The matrix $\mathbf{W}^{(0)}$, required for computing the estimates, now equals

$$\mathbf{W}^{(0)} = \begin{bmatrix} \mathbf{W}_{\mathbf{w}}^{(0)} & 0 \\ 0 & \mathbf{W}_{\mathbf{s}}^{(0)} \end{bmatrix},$$

where

$$\mathbf{W}_{\mathbf{w}}^{(0)} = \begin{bmatrix} \frac{p'_{\lambda_w}(|\beta_1^{(0)}|)}{|\beta_1^{(0)}|} & 0 & \dots & 0 \\ 0 & \frac{p'_{\lambda_w}(|\beta_2^{(0)}|)}{|\beta_2^{(0)}|} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{p'_{\lambda_w}(|\beta_{d_w^*}^{(0)}|)}{|\beta_{d_w^*}^{(0)}|} \end{bmatrix}$$

corresponds to the whole-plot effects, and

$$\mathbf{W}_{\mathbf{s}}^{(0)} = n \begin{bmatrix} \frac{p'_{\lambda_s}(|\beta_{d_w^*+1}^{(0)}|)}{|\beta_{d_w^*+1}^{(0)}|} & 0 & \dots & 0 \\ 0 & \frac{p'_{\lambda_s}(|\beta_{d_w^*+2}^{(0)}|)}{|\beta_{d_w^*+2}^{(0)}|} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{p'_{\lambda_s}(|\beta_{d^*}^{(0)}|)}{|\beta_{d^*}^{(0)}|} \end{bmatrix},$$

corresponds to the sub-plot effects. We denote by d^* the total number of nonzero model coefficients and by d_w^* the number of nonzero model coefficients that correspond to whole-plot effects.

As the penalized generalized least squares approach involves two different penalties for split-plot experiments, we had to estimate two thresholding parameters, λ_w and λ_s. We did so using generalized cross-validation. The resulting value for λ_w was larger than than for λ_s, which entails a larger penalty for the whole-plot effects. Again, at each step of the estimation procedure, the variance components σ_γ^2 and σ_ε^2 are re-estimated using REML.

4 Practical examples

This section is a proof-of-concept section in which we demonstrate the usefulness of penalized generalized least squares in a variety of experimental scenarios. We start with the simultaneous selection and estimation of models for a non-orthogonally blocked response surface experiment involving various responses, using Equation (5). Next, we focus on two different split-plot experiments, for which we use Equation (7). The first split-plot experiment we consider is a response surface experiment for which the ordinary least squares (OLS) estimator and the generalized least squares (GLS) estimator are equivalent. The split-plot experiment is a response surface experiment for which there is no equivalence between OLS and GLS. All of the data sets, each of which is in the public domain, involve quantitative experimental variables only and allow at least some of the quadratic effects to be estimated.

4.1 The pastry dough experiment

The first data set we study involves a non-orthogonally blocked response surface design and four different responses. The construction method for the design is described in Trinca and Gilmour (2000), while the different responses that we study are described in Gilmour and Ringrose (1999) and Gilmour and Trinca (2000). The experiment was aimed at improving the quality of a pastry dough. It involved seven blocks or four runs, and three experimental variables: flow rate (x_1), moisture content (x_2) and screw speed (x_3). The four responses in the data set are a longitudinal expansion index (y_1), a cross-sectional expansion index (y_2), and two measures of light transmission in two different bands of the spectrum (y_3 and y_4).

The estimates for the ten parameters of the full quadratic model for each of the four responses are displayed in Table 1. For each response, the table contains the estimates obtained using three versions of the penalized least squares method based on equation (5) as well as the estimates obtained using (non-penalized) generalized least squares and backward model selection. The key result is that each of the four methods results in the same final model for every response. In terms of variable selection, the three variants of the penalized least squares approach thus do equally well as manual backward model selection. The point estimates obtained using the hard thresholding penalty function are almost identical to those produced by (non-penalized) generalized least squares. The point estimates for LASSO (based on the L_1 penalty function) are those that differ most from the (non-penalized) generalized least squares estimates. The LASSO estimates are biased toward zero, and their standard errors are smaller than those for the three other methods.

4.2 The wind tunnel experiment

The second data set also involves four different responses (coefficient of lift — front (y_1), coefficient of lift — rear (y_2), drag (y_3) and lift over drag ratio (y_4)) and comes from a wind tunnel experiment carried out using a split-plot design. The design had nine whole plots of five runs each. Four experimental variables were studied: front ride height, rear ride height, yaw angle and grille coverage. The first two of these are whole-plot variables, whereas the others are sub-plot variables. Therefore, we denote the four experimental variables by w_1, w_2, s_1 and s_2, respectively. A special feature of the design was that only one of the quadratic whole-plot effects and only one of the quadratic sub-plot effects could be estimated. The data from the wind tunnel experiment are extensively described in Simpson, Kowalski and Landman (2004).

The estimates for the 15 parameters of the full quadratic model obtained using the three variants of the penalized least squares method and Equation (7) as well as using the non-penalized approach for each of the four responses are displayed in Table 2. Again, each of the four methods results in the same final model for every response. Moreover, the point estimates as well as the standard errors are nearly identical for all methods. It is striking that the differences between the methods are smaller for the data from the wind tunnel experiment. This is due to the small variance in each of the responses and to the experimental design used for the wind tunnel experiment, which is a crossed design with five runs per whole plot and only two variables that were varied within the whole plots, from which most of the model parameters can be estimated independently. This was not the case for the pastry dough experiment, which had only four runs per block and three variables that were varied within the blocks, so that most parameters could not be estimated independently.

4.3 The freeze-dried coffee experiment

The final data set that we study is from an experiment carried out to study the effect of several process variables on the retention of volatile compounds in the freeze drying of coffee. The setup and analysis of the freeze-dried coffee experiment were reported in Gilmour *et al.* (2000) and Gilmour and Goos (2009). Five variables were studied at three levels: the pressure in the drying chamber (w), the heating temperature (s_1), the initial solids content in the coffee solution (s_2), the thickness of the slab freeze dried as a batch (s_3), and the freezing rate (s_4). The experiment was conducted using a split-plot design with six whole plots, each containing five sub-plots. The pressure variable was a whole-plot variable, while the other four variables were sub-plot variables. Four of the responses measured were the amounts of the volatile compounds methylpirazine (y_1), benzaldehyde (y_2), 4-ethylbenzaldehyde (y_3) and 2-methoxy-4-methylphenol (y_4) retained after freeze-drying.

Table 1: Estimated coefficients and standard errors (in parentheses) for the pastry dough experiment

	Method	Int.	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	x_1^2	x_2^2	x_3^2
y_1	SCAD	11.8464	0.9944	−1.4556	0.5544	0	0	0	0	2.0222	0
		(0.4928)	(0.2351)	(0.2351)	(0.1725)	(−)	(−)	(−)	(−)	(0.4330)	(−)
	HARD	11.8464	0.9944	−1.4556	0.7556	0	0	0	0	2.0222	0
		(0.4928)	(0.2351)	(0.2351)	(0.2351)	(−)	(−)	(−)	(−)	(0.4330)	(−)
	LASSO	11.6229	0.8935	−1.3546	0.6546	0	0	0	0	1.8999	0
		(0.4536)	(0.2112)	(0.2188)	(0.2037)	(−)	(−)	(−)	(−)	(0.3632)	(−)
	GLS	11.8464	0.9944	−1.4556	0.7556	0	0	0	0	2.0222	0
		(0.4968)	(0.2351)	(0.2351)	(0.2351)	(−)	(−)	(−)	(−)	(0.4440)	(−)
y_2	SCAD	4.4655	0	−0.6233	0.3139	0	0	0	0	0.6871	0
		(0.1694)	(−)	(0.0842)	(0.0812)	(−)	(−)	(−)	(−)	(0.1546)	(−)
	HARD	4.4655	0	−0.6233	0.3256	0	0	0	0	0.6871	0
		(0.1694)	(−)	(0.0842)	(0.0842)	(−)	(−)	(−)	(−)	(0.1546)	(−)
	LASSO	4.4354	0	−0.6073	0.3096	0	0	0	0	0.6678	0
		(0.1628)	(−)	(0.0821)	(0.0801)	(−)	(−)	(−)	(−)	(0.1424)	(−)
	GLS	4.4655	0	−0.6233	0.3256	0	0	0	0	0.6870	0
		(0.1710)	(−)	(0.0842)	(0.0842)	(−)	(−)	(−)	(−)	(0.1587)	(−)
y_3	SCAD	13.1249	−0.1715	0.8783	−0.7094	0	0	0	0	−0.5904	0
		(0.3633)	(0.0765)	(0.0845)	(0.0845)	(−)	(−)	(−)	(−)	(0.1596)	(−)
	HARD	13.1249	−0.1894	0.8783	−0.7094	0	0	0	0	−0.5904	0
		(0.3633)	(0.0845)	(0.0845)	(0.0845)	(−)	(−)	(−)	(−)	(0.1596)	(−)
	LASSO	13.0271	−0.1847	0.8736	−0.7047	0	0	0	0	−0.5629	0
		(0.3601)	(0.0824)	(0.0841)	(0.0840)	(−)	(−)	(−)	(−)	(0.1549)	(−)
	GLS	13.1249	−0.1894	0.8783	−0.7094	0	0	0	0	−0.5904	0
		(0.3634)	(0.0846)	(0.0846)	(0.0846)	(−)	(−)	(−)	(−)	(0.1603)	(−)
y_4	SCAD	77.0784	0	−0.2767	0.2186	0	0	−0.3666	0	0	0.4180
		(0.1785)	(−)	(0.0700)	(0.0686)	(−)	(−)	(0.0838)	(−)	(−)	(0.1338)
	HARD	77.0781	0	−0.2844	0.2294	0	0	−0.3706	0	0	0.4185
		(0.1785)	(−)	(0.0720)	(0.0720)	(−)	(−)	(0.0848)	(−)	(−)	(0.1340)
	LASSO	77.0168	0	−0.2688	0.2138	0	0	−0.3490	0	0	0.3991
		(0.1736)	(−)	(0.0680)	(0.0671)	(−)	(−)	(0.0798)	(−)	(−)	(0.1179)
	GLS	77.0781	0	−0.2844	0.2294	0	0	−0.3706	0	0	0.4185
		(0.1792)	(−)	(0.0720)	(0.0720)	(−)	(−)	(0.0862)	(−)	(−)	(0.1363)

Table 2: Estimated coefficients and standard errors (in parentheses) for the wind tunnel experiment

	Method	Int.	w_1	w_2	s_1	s_2	w_1w_2	w_1s_1	w_1s_2	w_2s_1	w_2s_2	s_1s_2	w_1^2	s_1^2
	SCAD	−0.1226	0.0038	−0.0144	0.0067	−0.0255	−0.0046	−0.0018	0	0	0	0	0	0
		(0.0010)	(0.0011)	(0.0011)	(0.0005)	(0.0005)	(0.0011)	(0.0005)	(−)	(−)	(−)	(−)	(−)	(−)
	HARD	−0.1226	0.0038	−0.0144	0.0067	−0.0255	−0.0046	−0.0018	0	0	0	0	0	0
		(0.0010)	(0.0011)	(0.0011)	(0.0005)	(0.0005)	(0.0011)	(0.0005)	(−)	(−)	(−)	(−)	(−)	(−)
y_1	LASSO	−0.1226	0.0038	−0.0144	0.0067	−0.0255	−0.0046	−0.0018	0	0	0	0	0	0
		(0.0010)	(0.0011)	(0.0011)	(0.0005)	(0.0005)	(0.0011)	(0.0005)	(−)	(−)	(−)	(−)	(−)	(−)
	GLS	−0.1226	0.0039	−0.0144	0.0067	−0.0255	−0.0047	−0.0018	0	0	0	0	0	0
		(0.0012)	(0.0012)	(0.0012)	(0.0005)	(0.0005)	(0.0012)	(0.0005)	(−)	(−)	(−)	(−)	(−)	(−)
	SCAD	−0.2369	0.0120	−0.0146	0.0144	−0.0018	0	−0.0029	0.0020	0	−0.0025	0.0019	−0.0073	0.0046
		(0.0037)	(0.0012)	(0.0012)	(0.0007)	(0.0007)	(−)	(0.0008)	(0.0008)	(−)	(0.0008)	(0.0007)	(0.0037)	(0.0016)
	HARD	−0.2369	0.0120	−0.0146	0.0144	−0.0018	0	−0.0029	0.0020	0	−0.0025	0.0019	−0.0073	0.0046
		(0.0037)	(0.0012)	(0.0012)	(0.0007)	(0.0007)	(−)	(0.0008)	(0.0008)	(−)	(0.0008)	(0.0007)	(0.0037)	(0.0016)
y_2	LASSO	−0.2369	0.0120	−0.0146	0.0144	−0.0018	0	−0.0029	0.0020	0	−0.0025	0.0019	−0.0073	0.0046
		(0.0037)	(0.0012)	(0.0012)	(0.0007)	(0.0007)	(−)	(0.0008)	(0.0008)	(−)	(0.0008)	(0.0007)	(0.0037)	(0.0016)
	GLS	−0.2369	0.0120	−0.0146	0.0144	−0.0018	0	−0.0029	0.0020	0	−0.0025	0.0019	−0.0073	0.0046
		(0.0025)	(0.0014)	(0.0014)	(0.0008)	(0.0008)	(−)	(0.0009)	(0.0009)	(−)	(0.0009)	(0.0008)	(0.0022)	(0.0017)
	SCAD	0.4012	0.0086	0.0090	−0.0117	−0.0049	0	0	0	0	0.0008	−0.0006	0	0.0012
		(0.0007)	(0.0006)	(0.0006)	(0.0003)	(0.0003)	(−)	(−)	(−)	(−)	(0.0003)	(0.0003)	(−)	(0.0006)
	HARD	0.4012	0.0086	0.0090	−0.0117	−0.0049	0	0	0	0	0.0008	−0.0006	0	0.0012
		(0.0007)	(0.0006)	(0.0006)	(0.0003)	(0.0003)	(−)	(−)	(−)	(−)	(0.0003)	(0.0003)	(−)	(0.0006)
y_3	LASSO	0.4012	0.0086	0.0090	−0.0117	−0.0049	0	0	0	0	0.0008	−0.0006	0	0.0012
		(0.0007)	(0.0006)	(0.0006)	(0.0003)	(0.0003)	(−)	(−)	(−)	(−)	(0.0003)	(0.0003)	(−)	(0.0006)
	GLS	0.4012	0.0086	0.0090	−0.0117	−0.0049	0	0	0	0	0.0008	−0.0006	−	0.0012
		(0.0006)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(−)	(−)	(−)	(−)	(0.0003)	(0.0003)	(−)	(0.0006)
	SCAD	0.9010	−0.0586	0.0518	−0.0264	0.0792	0.0087	0.0096	−0.0071	0	0.0046	0	0	0
		(0.0023)	(0.0024)	(0.0024)	(0.0021)	(0.0021)	(0.0024)	(0.0023)	(0.0023)	(−)	(0.0023)	(−)	(−)	(−)
	HARD	0.9010	−0.0586	0.0518	−0.0264	0.0792	0.0087	0.0096	−0.0071	0	0.0046	0	0	0
		(0.0023)	(0.0024)	(0.0024)	(0.0021)	(0.0021)	(0.0024)	(0.0023)	(0.0023)	(−)	(0.0023)	(−)	(−)	(−)
y_4	LASSO	0.9010	−0.0586	0.0518	−0.0264	0.0792	0.0087	0.0096	−0.0071	0	0.0046	0	0	0
		(0.0023)	(0.0024)	(0.0024)	(0.0021)	(0.0021)	(0.0024)	(0.0023)	(0.0023)	(−)	(0.0023)	(−)	(−)	(−)
	GLS	0.9010	−0.0586	0.0518	−0.0264	0.0792	0.0087	0.0096	−0.0071	0	0.0046	0	0	0
		(0.0023)	(0.0024)	(0.0024)	(0.0021)	(0.0021)	(0.0024)	(0.0023)	(0.0023)	(−)	(0.0023)	(−)	(−)	(−)

Table 3: Estimated coefficients and standard errors (in parentheses) for the freeze-dried coffee experiment

	Method	Int.	w_1	s_1	s_2	s_3	s_4	w_1s_1	w_1s_2	w_1s_3	w_1s_4	s_1s_2	s_1s_3	s_1s_4	s_2s_3	s_2s_4	s_3s_4	w_1^2	s_1^2	s_2^2	s_3^2	s_4^2
	SCAD	0.1300	0	0	0	0	0	0	0.0774	0	0	0	0	0.0464	0	0	0	0	0	0	−0.0692	0
		(0.0334)	(−)	(−)	(−)	(−)	(−)	(−)	(0.0318)	(−)	(−)	(−)	(−)	(0.0277)	(−)	(−)	(−)	(−)	(−)	(−)	(0.0379)	(−)
	HARD	0.1303	0	0	0	0	0	0	0.0778	0	0	0	0	0.0469	0	0	0	0	0	0	−0.0696	0
		(0.0335)	(−)	(−)	(−)	(−)	(−)	(−)	(0.0319)	(−)	(−)	(−)	(−)	(0.0280)	(−)	(−)	(−)	(−)	(−)	(−)	(0.0381)	(−)
y_1	LASSO	0.1286	0	0	0	0	0	0	0.0767	0	0	0	0	0.0460	0	0	0	0	0	0	−0.0677	0
		(0.0329)	(−)	(−)	(−)	(−)	(−)	(−)	(0.0315)	(−)	(−)	(−)	(−)	(0.0276)	(−)	(−)	(−)	(−)	(−)	(−)	(0.0373)	(−)
	GLS	0.0885	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		(0.0278)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)
	SCAD	2.5376	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		(0.1738)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)
	HARD	2.5376	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		(0.1738)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)
y_2	LASSO	2.4785	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		(0.1697)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)
	GLS	2.5376	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		(0.1738)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)
	SCAD	1.9366	0	0	0	0	0	0	0	0	0	0.2342	0	0	0	0	0	0	0	0	0	0
		(0.0853)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(0.0930)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)
	HARD	1.9366	0	0	0	0	0	0	0	0	0	0.2519	0	0	0	0	0	0	0	0	0	0
		(0.0853)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(0.1000)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)
y_3	LASSO	1.9122	0	0	0	0	0	0	0	0	0	0.2183	0	0	0	0	0	0	0	0	0	0
		(0.0842)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(0.0867)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)
	GLS	1.9366	0	0	0	0	0	0	0	0	0	0.2520	0	0	0	0	0	0	0	0	0	0
		(0.0853)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(0.1000)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)
	SCAD	1.5050	0	0	−0.0702	0	0	0	0	0	0	0.4422	0	0	0	0	0	0	0	−0.1982	0.4847	0
		(0.0790)	(−)	(−)	(0.0550)	(−)	(−)	(−)	(−)	(−)	(−)	(0.0638)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(0.1327)	(0.1353)	(−)
	HARD	1.5063	0	0	−0.0763	0	0	0	0	0	0	0.4422	0	0	0	0	0	0	0	−0.2058	0.4902	0
		(0.0793)	(−)	(−)	(0.0598)	(−)	(−)	(−)	(−)	(−)	(−)	(0.0638)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(0.1379)	(0.1379)	(−)
y_4	LASSO	1.4690	0	0	−0.0704	0	0	0	0	0	0	0.4336	0	0	0	0	0	0	0	0	0.3388	0
		(0.0730)	(−)	(−)	(0.0559)	(−)	(−)	(−)	(−)	(−)	(−)	(0.0645)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(0.0924)	(−)
	GLS	1.4690	0	0	0	0	0	0	0	0	0	0.4422	0	0	0	0	0	0	0	0	0.3465	0
		(0.0820)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(0.0665)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(−)	(0.0924)	(−)

The estimates for the 21 parameters of the full quadratic models for the four responses are displayed in Table 3. For two of the four responses, y_2 and y_3, the three variants of the penalized least squares approach and the (non-penalized) generalized least squares approach result in the same set of selected variables. For these responses, using the L_1 penalty function (LASSO) yields point estimates that are shrunk to zero to a substantial extent and smaller standard errors too. The three other approach yield very similar point estimates and standard errors. Compared to the results for the wind tunnel experiment, the differences between the methods are larger in the freeze-dried coffee example. Comparing Tables 1 and 3 shows that the differences between the methods in the freeze-dried coffee example are similar to those in the pastry dough experiment. The reason is that the experimental design used in the freeze-dried coffee experiment was not as orthogonal as the design used in the wind tunnel experiment, because four variables were varied within the blocks.

For the response y_1, there is a substantial difference between the penalized least squares approaches and the (non-penalized) generalized least squares. The penalized least squares approaches all retain three terms in the model, whereas the non-penalized approach leads to the selection of the null model. However, when using the non-penalized generalized least squares approach, there are various parameter estimates that are very nearly significant at the 5% level. Two of these are the interaction effect between pressure (w) and initial solids content (s_2) and the quadratic effect of thickness (s_3), which are present in the model selected by the penalized least squares method. Hence, the difference between the (non-penalized) least squares approach and the penalized least squares variants is not dramatic. Similar observations can be made for the response y_4. The penalized least squares approach retains four terms in the model when the SCAD and HARD penalty functions are used and three terms when the LASSO approach, involving the L_1 penalty function, is used. The non-penalized approach selects only two terms. However, the terms that are not included in the non-penalized approach are nearly significant at the 5% level. As a result, also for the response y_4, the difference between the penalized and non-penalized approach is not dramatic.

Of the three penalty functions, the L_1 penalty function (LASSO) leads to the poorest results in the sense that its estimates are shrunk toward zero to a substantial extent for non-orthogonal designs. In these cases, the standard errors resulting from the L_1 penalty function are also smaller than for the other methods. Therefore, we do not include the L_1 penalty function (LASSO) in our simulation study.

5 Simulation Study

From the practical examples, it is clear that the penalized generalized least squares approach, when utilized with the SCAD or the hard thresholding penalty function, is at the very least a promising approach for automated model selection for data from blocked experiments and split-plot experiments. In this section, we study the properties of the penalized generalized least squares approach in detail by means of a simulation study. The goals of the simulation study are (i) to investigate its performance under the null model in which the experimental variables have zero main effects, interaction effects and quadratic effects, and (ii) to investigate to what extent the penalized generalized least squares is capable of detecting nonzero effects of the experimental variables. As a benchmark for the performance of the penalized generalized least squares approach, we utilize the (non-penalized) generalized least squares analysis using REML estimates of the variance components σ_γ^2 and σ_ε^2, as recommended by Gilmour and Trinca (2000) and Letsinger *et al.* (1996). In the discussion below, we label this approach REML-GLS.

5.1 Setup

In our simulation study, we adopt an approach similar to that of Goos *et al.* (2006), who studied the type I and type II error rates of various split-plot response surface designs with a traditional generalized least squares analysis using a significance level of 5%, where the variance components are estimated using REML. Like Goos *et al.* (2006), we use a backward selection approach for the generalized least squares approach, where we fit the full model and drop all the non-significant terms in one step to arrive at the final model.

First, we study the performance of the penalized generalized least squares approaches under the null model. This allows us to verify to what extent the approaches control the type I error rate. Next, we study the type II error rate of the penalized generalized least squares approaches. In our simulation study, we set $\sigma_\gamma^2 + \sigma_\varepsilon^2 = 20$, without loss of generality, and used three variance ratios, $\eta = 1/10$, 1 and 10. Similar values for η have been obtained from the analysis of data from many blocked and split-plot experiments (see, for instance, Letsinger *et al.* (1996), Littell *et al.* (1996), Gilmour and Trinca (2000) and Webb *et al.* (2004)). We compare our results to those of Goos *et al.* (2006), who showed that the type I error rate for whole-plot factor effects and quadratic sub-plot effects are hard to control if the number of whole plots of a split-plot response surface design is small. According to their results, only the combined REML-GLS analysis with degrees of freedom obtained from the method of Kenward and Roger (1997) seems capable of controlling the type I error for most of the effects, with some exceptions for the whole-plot effects and quadratic effects of the sub-plot factors. Other degrees of freedom methods have substantially too large a type I error rate in most circumstances.

Some of the the type II error rates obtained by Goos *et al.* (2006) with the REML-GLS approach and Kenward-Roger degrees of freedom are substantial. According to their results, even whole-plot effects as large as the standard deviation of the responses remain undetected in at least 60% of the cases. For the sub-plot effects, the type II error rates are quite small, except when they are not orthogonal to the whole plots. Because the variables in a blocked experiment are treated identically as sub-plot variables in a split-plot design, the type II error rates for all the effects in a blocked experiment should be rather small.

5.2 Type I error rates

For each of the designs utilized in Section 4, we simulated 1000 data sets using a null model with $\eta = 1/10$, 1 and 10 for $\sigma^2 = \sigma_\gamma^2 + \sigma_\varepsilon^2 = 20$, and analyzed each data set with the penalized generalized least squares approach. This allows us to study the type I error rate that the method yields. We mainly report results obtained using the SCAD penalty function because, in general, that penalty function leads to substantially smaller type I error rates than the hard thresholding penalty function.

5.2.1 Pastry dough experiment

The type I error rates for the design for the pastry dough experiment obtained using the SCAD and hard thresholding penalty functions are given in Table 4. For the main effects and the two-factor interaction effects, the type I error rates for the SCAD penalty function lie between 0.004 and 0.038, which is acceptable. The type I error rates for the quadratic effects are noticeably larger and range from 0.036 to 0.170. The qualitatively different behavior of the quadratic effects is explained by the fact that the quadratic effects are not nearly as orthogonal to the blocks as the main effects and the two-factor interaction effects.

A clear pattern in the type I error rates when using the SCAD penalty function is that they decrease with η for each of the effects. This is because large values of η correspond with small values of σ_ε^2, so that the standard errors for each of the factor effects are substantially smaller if η is large. As a matter of fact, in a blocked experiment, it is mainly the residual error variance σ_ε^2 that determines the absolute magnitude of the parameter estimates' standard errors. For the hard thresholding penalty function the type I error rates are unacceptably large.

For the main effects and the two-factor interaction effects, the type I error rates for the non-penalized generalized least squares lie between 0.041 and 0.066, which is in line with the significance level chosen. The type I error rates for the quadratic effects range from 0.048 to 0.079. They are noticeably smaller than those of the SCAD method, for the cases with eta equal to 0.1 and 1, and much smaller than those for the HARD penalty function. The type I error rates for the quadratic effects are close to 5% when the non-penalized generalized least squares approach is used, independently of the variance ratio η.

Table 4: Type I error rates for the pastry dough experiment

Method	η	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	x_1^2	x_2^2	x_3^2
SCAD	0.1	0.018	0.017	0.011	0.031	0.032	0.038	0.148	0.170	0.139
	1	0.011	0.008	0.019	0.026	0.034	0.018	0.085	0.088	0.083
	10	0.009	0.009	0.010	0.008	0.004	0.016	0.036	0.043	0.041
HARD	0.1	0.161	0.172	0.164	0.199	0.173	0.202	0.284	0.300	0.275
	1	0.090	0.071	0.082	0.145	0.147	0.138	0.274	0.301	0.286
	10	0.040	0.045	0.046	0.073	0.075	0.089	0.303	0.316	0.320
GLS	0.1	0.046	0.053	0.046	0.051	0.059	0.066	0.065	0.079	0.057
	1	0.046	0.051	0.048	0.053	0.062	0.041	0.055	0.048	0.051
	10	0.057	0.054	0.061	0.036	0.045	0.056	0.050	0.066	0.064

5.2.2 Wind tunnel experiment

The type I error rates for the wind tunnel experiment obtained using the SCAD penalty function are given in Table 5. Because of the poor performance of the hard thresholding penalty function, we do not report any results for this method.

The penalized generalized least squares approach with the SCAD penalty function keeps the type I error rate under control very well for the main effects, the quadratic effect and the interaction effect of the two whole-plot factors w_1 and w_2 for all η values. The type I error rates lie between 0.0% and 2.9% for these effects. For the main effects of the sub-plot factors and the interaction effects involving the sub-plot factors, the type I error rates for the SCAD method lie between 0.1% and 1.5%. So, the type I error rate is also under control for these effects. This is also the case for the quadratic effect of the sub-plot variable s_1, except for the case where η is 0.1.

The type I error rates for the non-penalized generalized least squares lie between 4.1% and 6.4% for the whole-plot effects, whereas those for the effects that involve sub-plot factors range from 3.1% to 6.5%. Hence, here too, the type I error rate is close to the 5% significance level chosen.

5.2.3 Freeze-dried coffee experiment

The final results on type I error rates are those reported in Table 6 for the freeze-dried coffee experiment. The design for the freeze-dried coffee experiment involved fewer whole plots than the wind tunnel experiment. The only visible impact of this is a slightly higher type I error rate for all quadratic sub-plot effects for the penalized method. For the whole-plot effects, the penalized generalized least squares approach with the SCAD penalty function yields a smaller type I error rate than the traditional method without the penalty. This is also the case for the main effects and interaction effects of the sub-plot variables. For the interaction effects between whole-plot and sub-plot factors, the opposite is true.

5.3 Type II error rates

For each design used in Section 4, we also simulated 1000 data sets using a non-null model with $\eta = 1/10$, 1 and 10 for $\sigma^2 = \sigma_\gamma^2 + \sigma_\varepsilon^2 = 20$. When simulating data, we ensured that the largest effects were of the same order of magnitude as σ. This is the typical effect size that an experimenter certainly wants to detect. We analyzed the simulated data using the penalized generalized least squares approach with the SCAD penalty function and recorded the type II error rates.

5.3.1 Pastry dough experiment

The type II error rates obtained using the SCAD penalty function or using the traditional generalized least squares approach for the design for the pastry dough experiment are given in Table 7. These results are based on the model

$$E(Y) = 50 + 4x_1 + 2x_2 + x_3 - 4x_1x_2 - 2x_1x_3 - x_2x_3 + 4x_1^2 + 2x_2^2 + x_3^2,$$

so that there is a large, a medium-sized and a small main effect, interaction effect and quadratic effect to detect.

The type II error rates for the penalized method are very low for η equal to 1 or 10 for the largest main effect and for the largest interaction effect, i.e. for the main effect and the interaction effect that are as large as the standard deviation of the responses, σ. For the quadratic effect of the same size, the type II error rate is negligible for $\eta = 10$ but as large as 32.9% for $\eta = 0.1$. The medium-sized main effect and interaction effects have smaller type II error rates when we use generalized least squares without penalization. The results also show that effects of one quarter the size of σ are hard to detect and have large type II error rates, and that the power for detecting medium-size and small quadratic effects is limited for any value of η for the penalized generalized least squares as well as the non-penalized generalized least squared. For the quadratic effects, however, the performance of the penalized approach is better than that of the non-penalized one.

Table 5: Type I error rates for the wind tunnel experiment

Method	η	w_1	w_2	s_1	s_2	w_1w_2	w_1s_1	w_1s_2	w_2s_1	w_2s_2	s_1s_2	w_1^2	s_1^2
SCAD	0.1	0.007	0.005	0.015	0.008	0.006	0.010	0.007	0.014	0.007	0.008	0.000	0.152
	1	0.029	0.018	0.004	0.007	0.014	0.008	0.014	0.003	0.008	0.008	0.006	0.066
	10	0.014	0.014	0.005	0.006	0.010	0.003	0.004	0.001	0.003	0.003	0.006	0.014
GLS	0.1	0.064	0.052	0.031	0.038	0.057	0.049	0.057	0.055	0.042	0.051	0.058	0.036
	1	0.044	0.047	0.060	0.057	0.048	0.065	0.045	0.063	0.055	0.043	0.054	0.043
	10	0.041	0.042	0.052	0.065	0.048	0.052	0.049	0.046	0.059	0.047	0.050	0.047

Table 6: Type I error rates for the freeze-dried coffee experiment

Method	η	w_1	s_1	s_2	s_3	s_4	w_1s_1	w_1s_2	w_1s_3	w_1s_4	s_1s_2	s_1s_3	s_1s_4	s_2s_3	s_2s_4	s_3s_4	w_1^2	s_1^2	s_2^2	s_3^2	s_4^2
SCAD	0.1	0.011	0.039	0.045	0.040	0.054	0.105	0.104	0.105	0.111	0.040	0.046	0.055	0.050	0.041	0.048	0.003	0.220	0.208	0.215	0.227
	1	0.027	0.024	0.028	0.022	0.032	0.073	0.076	0.066	0.088	0.030	0.030	0.052	0.029	0.031	0.049	0.018	0.153	0.136	0.135	0.167
	10	0.037	0.031	0.031	0.035	0.030	0.036	0.031	0.031	0.024	0.025	0.027	0.042	0.029	0.025	0.034	0.027	0.079	0.080	0.074	0.084
GLS	0.1	0.047	0.063	0.055	0.058	0.070	0.056	0.064	0.065	0.064	0.059	0.064	0.055	0.065	0.061	0.046	0.035	0.062	0.057	0.066	0.070
	1	0.059	0.052	0.049	0.046	0.065	0.056	0.068	0.071	0.072	0.047	0.056	0.057	0.058	0.050	0.052	0.054	0.058	0.048	0.055	0.064
	10	0.055	0.063	0.047	0.068	0.059	0.046	0.047	0.052	0.052	0.049	0.057	0.056	0.066	0.050	0.043	0.040	0.059	0.058	0.047	0.060

Table 7: Type II error rates for the pastry dough experiment, along with the model coefficients used in the simulations

		x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	x_1^2	x_2^2	x_3^2
Method	η	4	2	1	-4	-2	-1	4	2	1
SCAD	0.1	0.092	0.685	0.914	0.119	0.686	0.892	0.329	0.611	0.706
	1	0.014	0.533	0.895	0.039	0.587	0.878	0.244	0.609	0.743
	10	0.000	0.129	0.761	0.000	0.193	0.765	0.003	0.335	0.749
GLS	0.1	0.032	0.507	0.821	0.080	0.595	0.838	0.693	0.870	0.921
	1	0.002	0.300	0.773	0.012	0.446	0.809	0.511	0.845	0.921
	10	0.000	0.000	0.165	0.000	0.000	0.331	0.003	0.365	0.778

5.3.2 Wind tunnel experiment

The type II error rates for the wind tunnel experiment obtained using the SCAD penalty function and the non-penalized approach are displayed in Table 8. We generated data using the model

$$E(Y) = 50 + 4w_1 + 2w_2 + 4s_1 + 2s_2 + 4w_1w_2 + w_1s_1 + 4w_1s_2 + 2w_2s_1 \\ + 6w_2s_2 - 2s_1s_2 + 4w_1^2 - 4s_1^2,$$

Note that this model only involves two quadratic terms because the design actually run did not allow for the estimation of all the quadratic effects. It contains large, medium-sized and small effects. Large main effects and interaction effects of the sub-plot factors s_1 and s_2 are detected almost with certainty for each of the three η values in our study, by the penalized as well as the non-penalized approach. For medium-sized sub-plot main effects and interaction effects, the type II error rate is almost zero when $\eta = 10$ and the penalized approach is used, but the type II error rate is substantially larger for low values of η. In some cases, the type II error rates are much better for the penalized approach than for the non-penalized one, but in others, it is the other way around. The type II error rate for the main effects, the interaction effects and the quadratic effect of the whole-plot factors for the penalized method are large compared to that for the non-penalized method.

5.3.3 Freeze-dried coffee experiment

The results on type II error rates for the freeze-dried coffee experiment are reported in Table 9. These results were obtained using the model

$$E(Y) = 50 + 4w + 4s_1 - 3s_2 + 2s_3 - s_4 - 4ws_1 + 3ws_2 - ws_3 + ws_4 + 4s_1s_2 \\ + 3s_1s_3 + 2s_1s_4 - s_2s_3 + 4w^2 + 4s_1^2 - 3s_2^2 + 2s_2^3 - s_2^4$$

for generating the data. The results for the freeze-dried coffee experiment are completely in line with those for the wind tunnel experiment. The type II error rate for the main and the quadratic effect of the whole-plot variable w is worse if we apply penalized least squares. This is true as well for the main effects of the sub-plot factors and the interaction effects involving sub-plot factors only. On the contrary, the type II error rates for interaction effects involving a whole-plot factor and a sub-plot factor are smaller for the penalized method than for the non-penalized method when η is small, but larger when η is large. The type II errors for the quadratic sub-plot effects exhibit a similar pattern.

6 Discussion

In this article, we have shown how to adapt the penalized least squares approach so that it can cope with correlated observations. We have done so by focusing on industrial experiments run in blocks and run in a split-plot format. We have named the newly developed approach penalized generalized least squares. One version of the new approach utilizes equally sized penalties for all factor effects and is suitable for experiments run in blocks. A second version of the approach uses larger penalties for the whole-plot factor effects than for the sub-plot factor effects.

We have studied three different sorts of penalty functions for use in the two versions of the penalized generalized least squares approach. Our analysis of data from various real-life experiments showed that the SCAD penalty function and the hard thresholding penalty function performed similarly and were superior to the L_1 penalty function (which is known as LASSO). In a subsequent simulation study, it turned out that the SCAD penalty function controlled the type I error substantially better than the hard thresholding penalty function. Hence, we recommend utilizing the SCAD penalty function in the penalized generalized least squares approach. This recommendation is in line with existing literature that deals with uncorrelated observations.

Both our analysis of the real-life data sets and the simulation study demonstrate that the penalized generalized least squares approach does not only do a good job when it comes to identifying non-zero effects, but it also estimates the effects as well as the conventional generalized least squares approach. What also became clear from our simulation study is that the penalized generalized least squares approach allows the type I error to be controlled well. For the whole-plot effects in split-plot models, this goes at the expense of the type II errors. For other factor effects, except for quadratic effects in blocked experiments and quadratic sub-plot effects in split-plot experiments when η is as small as 0.1, the type I error rates we obtained for the penalized generalized least squares approach are small. For the type II error rate, the results are mixed, with the penalized generalized least squares approach being better than the non-penalized approach in some cases and worse in other cases.

Table 8: Type II error rates for the wind tunnel experiment, along with the model coefficients used in the simulations

		w_1	w_2	s_1	s_2	w_1w_2	w_1s_1	w_1s_2	w_2s_1	w_2s_2	s_1s_2	w_1^2	s_1^2
Method	η	4	2	4	2	4	1	4	2	6	−2	4	−4
SCAD	0.1	0.889	0.948	0.005	0.553	0.889	0.916	0.010	0.535	0.000	0.539	0.999	0.210
	1	0.915	0.949	0.004	0.434	0.911	0.928	0.005	0.477	0.000	0.406	0.991	0.139
	10	0.954	0.969	0.000	0.012	0.954	0.655	0.000	0.021	0.000	0.011	0.991	0.000
GLS	0.1	0.000	0.216	0.066	0.538	0.220	0.756	0.230	0.000	0.000	0.055	0.300	0.770
	1	0.000	0.042	0.313	0.745	0.040	0.590	0.062	0.000	0.000	0.303	0.082	0.863
	10	0.000	0.000	0.517	0.832	0.000	0.020	0.000	0.000	0.000	0.485	0.000	0.908

Table 9: Type II error rates for the freeze-dried coffee experiment, along with the model coefficients used in the simulations

		w_1	s_1	s_2	s_3	s_4	w_1s_1	w_1s_2	w_1s_3	w_1s_4	s_1s_2	s_1s_3	s_1s_4	s_2s_3	w_1^2	s_1^2	s_2^2	s_3^2	s_4^2
Method	η	4	4	−3	2	−1	−4	3	−1	1	4	3	2	−1	4	4	−3	2	−1
SCAD	0.1	0.928	0.177	0.400	0.728	0.913	0.355	0.563	0.791	0.765	0.227	0.404	0.742	0.906	0.977	0.624	0.548	0.787	0.775
	1	0.919	0.087	0.292	0.668	0.931	0.347	0.548	0.816	0.772	0.110	0.322	0.734	0.933	0.971	0.545	0.541	0.784	0.807
	10	0.912	0.242	0.004	0.166	0.724	0.017	0.190	0.672	0.607	0.000	0.244	0.478	0.695	0.729	0.143	0.192	0.525	0.728
GLS	0.1	0.456	0.159	0.389	0.675	0.879	0.508	0.702	0.894	0.934	0.180	0.289	0.742	0.860	0.769	0.717	0.791	0.877	0.919
	1	0.646	0.044	0.230	0.519	0.852	0.374	0.627	0.901	0.937	0.051	0.105	0.688	0.855	0.828	0.597	0.742	0.846	0.922
	10	0.743	0.000	0.000	0.007	0.459	0.002	0.064	0.742	0.814	0.000	0.000	0.104	0.430	0.888	0.023	0.150	0.471	0.818

In our view, the penalized generalized least squares approach is a useful technique for analyzing data from blocked and split-plot experiments. The method is intended to overcome the criticisms that are raised against stepwise regression. The method circumvents the multiple testing issue and the difficulties in interpreting p values that are conditional upon what terms have been dropped from and added to the model in the stepwise regression process. Another advantage of the approach is that model selection and model estimation are done in a single step. This implies that the standard errors resulting from the penalized generalized least squares approach are justified theoretically, whereas this is not the case for standard errors resulting from stepwise regression.

The fact that model estimation and selection are done in one step also has a practical advantage over stepwise regression with correlated observations: no manual interventions are required from the researcher when using the penalized generalized least squares approach. As a matter of fact, with current software packages suitable for analyzing data from blocked and split-plot response surface experiments, the researcher has to perform several successive regression analyses manually. This is especially cumbersome for experiments that have multiple responses. As the examples in Section 4 show, the occurrence of multiple responses is the norm, rather than the exception. Obviously, an automated model selection and estimation tool is an important aid in the search for appropriate models for each of the responses. Our MATLAB implementation of the penalized generalized least squares approach is available from the first author.

We strongly believe that the penalized generalized least squares approach can benefit from further research. For example, it would be interesting to develop an approach that preserves model hierarchy. Also, the usefulness of other choices for the thresholding parameters based on the Akaike Information Criterion (AIC, Akaike 1974) and the Bayesian Information Criterion (BIC, Schwarz 1978) could be investigated. Similar work was already done in the literature on penalized least squares for uncorrelated responses (see, e.g., Wang *et al.* 2007 and Zhang *et al.* 2010. Another useful topic of research would be to study alternative estimation procedures (see, e.g., Zou and Li 2008, who use the LARS algorithm (Efron *et al.* 2004) to obtain penalized least squares estimates).

Acknowledgment

We are grateful to Runze Li and Dennis Lin for sending us the code of their implementation of the penalized least squares approach for uncorrelated observations, and to Steven Gilmour for providing us with the complete data set from the freeze-dried coffee experiment.

References

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control* **19**: 716–723.

Androulakis, E., Koukouvinos, C., Mylona, K. and Vonta, F. (2010). A real survival analysis application via variable selection methods for Cox's proportional hazards model, *Journal of Applied Statistics* **47**: 1399–1406.

Antoniadis, A. (1997). "Wavelets in statistics: a review (with discussion)", *Journal of the Italian Statistical Association* **6**: 97–144.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation, *Numerische Mathematik* **31**: 377–403.

Efron, B., Hastie, T., Johnstone, I. and Tibshirany, R. (2004). Least angle regression, *The Annals of Statistics* **32**: 407–499.

Fan, J. (1997). Comments on "Wavelets in statistics: a review" by A. Antoniadis, *Journal of the Italian Statistical Association* **6**: 131–138.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association* **96**: 1348–1360.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model, *The Annals of Statistics* **30**: 74–99.

Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery, *in* M. Sanz-Sole, J. Soria, V. J. L. and J. Verdera (eds), *Proceedings of the International Congress of Mathematicians, Vol. III*, pp. 595–622.

Gilmour, S. G. and Goos, P. (2009). Bayesian analysis of data from multi-stratum response surface designs, *Applied Statistics* **58**: 467–484.

Gilmour, S. G., Pardo, J. M., Trinca, L. A., Niranjan, K. and Mottram, D. S. (2000). A split-unit response surface design for improving aroma retention in freeze dried coffee, *Proceedings of the 6th European Conference on Food-Industry and Statistics*, pp. 18.0–18.9.

Gilmour, S. G. and Ringrose, T. J. (1999). Controlling processes in food technology by simplifying the canonical form of fitted response surfaces, *Applied Statistics* **48**: 91–101.

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.

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

Karagrigoriou, A., Koukouvinos, C. and Mylona, K. (2010). On the advantages of the nonconcave penalized likelihood model selection method with minimum prediction errors in large-scale medical studies, *Journal of Applied Statistics* **37**: 13–24.

Kenward, M. G. and Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood, *Biometrics* **53**: 983–997.

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

Li, R. and Lin, D. K. J. (2002). Data analysis in supersaturated designs., *Statistics & Probability Letters* **59**: 135–144.

Littell, R. C., Milliken, G. A., Stroup, W. W. and Wolfinger, R. D. (1996). *SAS System for Mixed Models*, Cary: SAS Institute.

Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics* **6**: 461–464.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992). *Variance Component Estimation*, New York: Academic Press.

Simpson, J. R., Kowalski, S. M. and Landman, D. (2004). Experimentation with randomization restrictions: Targeting practical implementation, *Quality and Reliability Engineering International* **20**: 481–495.

Tibshirany, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Ser. B* **58**: 267–288.

Trinca, L. A. and Gilmour, S. G. (2000). An algorithm for arranging response surface designs in small blocks, *Computational Statistics and Data Analysis* **33**: 25–43.

Wang, H., Li, R. and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method, *Biometrika* **94**: 553–568.

Webb, D., Lucas, J. M. and Borkowski, J. J. (2004). Factorial experiments when factor levels are not necessarily reset, *Journal of Quality Technology* **36**: 1–11.

Zhang, Y., Li, R. and Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion, *Journal of the American Statistical Association* **105**: 312–323.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models, *The Annals of Statistics* **36**: 1509–1533.