Greedy Variables Selection in Experimental Studies

Joachim Kunert and **Claus Weihs***
Fachbereich Statistik, Universität Dortmund, Germany
*weihs@statistik.uni-dortmund.de
May 2001

Abstract This paper derives optimal conditions for greedy variables selection, and discusses whether such conditions are met by standard experimental designs. On the other hand, it is shown that a well designed experiment might lead to a factor structure very appropriate for cross validation, namely a certain balance in the observations together with orthogonality of the factors. In contrast, however, it is shown that variables selection based on cross validation is not appropriate for saturated orthogonal designs. Moreover, it is demonstrated that otherwise optimal designs like rotatable designs may be sub-optimal for cross validation.

Keywords: variables selection, stepwise regression, cross validation, experimental design screening, optimisation

1. Introduction

Variables selection methods come along under different names in the literature. In the computer science and machine learning related literature the term 'feature selection' is standard (cp., e.g., Weiss and Kulikowski, 1991, as well as Weiss and Indurkhya, 1998). Also, there is a relation to the more general term 'model selection', which is, in particular, used in the framework of Bayesian statistics (cp., e.g., Gelfand, 1996). In this paper the term 'variables selection' stands for the use of cross validation to select 'important' influential variables for certain responses in observational as well as in experimental studies. Such procedures are very well standard in statistical software (see, e.g. SAS-STAT User's Guide, PROC REG and PROC RSREG). In this paper we derive optimal conditions for cross validation, and discuss them in particular in experimental studies.

In section 2 we introduce variables selection by means of 'greedy' stepwise forward selection (cp. Weihs, 1993, and Weihs and Jessenberger, 1999). Section 3 discusses the use of cross validation in experimental studies from a theoretical viewpoint and gives conditions favorable and unfavorable for variables selection by means of cross validation. In section 4 we discuss the optimal conditions in the framework of experimental design.

2. Greedy variables selection

In order to identify those **factors which mainly influence a target variable** in a linear model $y = X\beta + \varepsilon$, one can build the linear model for the target by, e.g., successively including the influential factors. This is done by identifying first that factor x_k with the biggest effect on the target y, then the factor with the biggest additional effect, etc. until no 'essential' improvement of model fit can be observed.

Such a method is called forward selection or **stepwise regression with forward selection** if the 'least squares criterion' is used to judge the effect size. If a factor once chosen is always kept in the model, the method is called **'greedy' forward selection**.

Even if there are **less observations** n **than possibly influential factors** K, then stepwise regression allows to study whether the target can be modelled adequately with $B < n-1$ factors.

Essential for the functionality of such a method is the choice of an appropriate measure for (the actually obtained) model adequacy, here expressed by the predictive power. One possibility for measuring predictive power is **cross validation**. A special version of cross validation is **leave one out**, where each observation of the sample is left out once individually, n regressions with $n-1$ observations each are carried out, and each of the left out observations is predicted, based on the coefficients estimated by the remaining observations. This way, for each observation one gets an oblique prediction. The differences of these to the true observations are used to define the predictive power.

Predictive power based on leave one out

Assuming that the current model of the target Y has K influential factors, the **predictive power** corresponding to a target Y is defined as

$R^2_{cv} := 1 - \dfrac{RSS_{cv}}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ (**R^2 cross validated**), where $RSS_{cv} := \sum_{i=1}^{n} \hat{v}_i^2$,

and the $\hat{v}_i := y_i - \hat{y}_{i,cv}$, $i = 1, ..., n$, are the prediction errors, y_i is the i^{th} observation of the target, $\hat{y}_{i,cv} := x_i^T \hat{\beta}(i)$ is the prediction of the i^{th} observation of the target, x_i^T is the i^{th} row of the design matrix, and $\hat{\beta}(i)$ the estimated coefficients vector based on all observations except the i^{th}.

In what follows R^2_{cv} will be used as the performance measure in greedy forward selection.

Stepwise regression with greedy forward selection

In **stepwise regression with greedy forward selection** first that factor is chosen out of the possibly influential factors, which maximises R^2_{cv} in a model with (possibly an) increment and one factor only. Then another factor is chosen, the addition of which to the model increases R^2_{cv} most, etc. until R^2_{cv} does not improve anymore.

Let us now discuss the mathematical background of this method, in particular optimal conditions for its application.

3. Cross validation

3.1 *The idea of cross validation*

Assume we have n $(K+L)$-dimensional vectors of observations (x_i, y_i), $i = 1, ..., n$. We assume that these are a random sample of independent identically distributed $(K+L)$-dimensional variables with unknown joint distribution $P^{X,Y}$.

Assume that for an $(n+1)^{st}$ observation we have as yet only observed the part x_{n+1} and we want to use this to make a good prediction for y_{n+1}. We assume that (x_{n+1},y_{n+1}) also has the distribution $P^{X,Y}$. The prediction $\hat{y}_{n+1}$ will be some function f of the observed x_{n+1}, $\hat{y}_{n+1} = f(x_{n+1})$, and the so-called **prediction rule** f will be determined by regressing y on x in the set $(x_1,y_1), ..., (x_n,y_n)$.

To see how well this prediction works, let us use **cross validation** with the leave one out technique which provides an unbiased estimate of $E(\left\|Y-\hat{Y}\right\|^2)$. We omit one observation (x_s,y_s) from the sample $(x_1,y_1), ..., (x_n,y_n)$, then we calculate the prediction formula from the remaining n-1 observations and we use this formula to calculate a prediction $\hat{y}_{s,cv}$ for the one observation that has been omitted. Then $\left\|y_s-\hat{y}_{s,cv}\right\|^2$ is an estimate of $E(\left\|Y-\hat{Y}\right\|^2)$. Repeating this for every s, we get the cross validated estimate $\frac{1}{n}\sum_{s=1}^{n}\left\|y_s-\hat{y}_{s,cv}\right\|^2$, which is equal to RSS_{cv} / n considered in section 2 in the case of one target Y only.

If we want to **compare several models** for prediction, then we might want to select the one for which $\frac{1}{n}\sum_{s=1}^{n}\left\|y_s-\hat{y}_{s,cv}\right\|^2$ is smallest. Note that this destroys the validity of $\frac{1}{n}\sum_{s=1}^{n}\left\|y_s-\hat{y}_{s,cv}\right\|^2$ as a predictor of $E(\left\|Y-\hat{Y}\right\|^2)$. To see this, assume there are several

models, all of which have basically the same predictive power, such that $\frac{1}{n}\sum_{s=1}^{n}\left\|y_s - \hat{y}_{s,cv}\right\|^2$ has basically the same distribution for each of the predictors. Consequently, since we select the one model with the minimal $\frac{1}{n}\sum_{s=1}^{n}\left\|y_s - \hat{y}_{s,cv}\right\|^2$, we underestimate $E(\left\|Y - \hat{Y}\right\|^2)$. Therefore, $\frac{1}{n}\sum_{s=1}^{n}\left\|y_s - \hat{y}_{s,cv}\right\|^2$ is a good means to select an appropriate model, but it then gives a too optimistic view of the performance of the model which is actually chosen.

3.2 *Cross validation and designed experiments*

The situation is very different, however, if a designed experiment is carried out. With a designed experiment, the set $\{x_1, ..., x_n\}$ is deliberately chosen. Therefore, the (x_i, y_i) are not identically distributed. Additionally, the new observation (x_{n+1}, y_{n+1}) is not just another observation with the same distribution, but the point x_{n+1} for which we want to predict y is a fixed point of interest. In most cases it is a point where we did not observe y during the experiment.

Example 1 (Simulation): We take an artificial example to show, how cross-validation can be misleading for a designed experiment. Assume that we have observed a one-dimensional x at points $x_1 = 0.1$, $x_2 = 0.2$, $x_3 = 0.3$, $x_4 = 0.4$, $x_5 = 0.5$, $x_6 = 0.6$, $x_7 = 10$. Further assume that the conditional distribution of a one-dimensional y given x is the normal distribution with expectation $10+x$ and variance 1.

If the data were derived from an observational study, then there is a good argument that the observation 7 might be an outlier. If the data were derived from an experiment, then we have designed x_7 to be different from the other x_i. Therefore there is no reason why this observation should be less reliable than the others.

Assuming the model described above, we simulated 10,000 data sets with the given x_i, i=1,...,7, and corresponding y_i. There are two simple models which we might use for prediction. The first model does not take account of the x,

(M1) $\quad y = \mu + \varepsilon,$

whereas the second model uses x,

(M2) $\quad y = \mu + \beta x + \varepsilon.$

We know from the way how we have simulated the data that (M2) is the correct model. If the data were observations from a true experiment or a true observational study, then we would not know which is the right model. We would have to decide from the data which model appears more adequate.

We compare two methods for deciding between the models. The first method is cross validation. For each of (M1) and (M2) we calculate $\frac{1}{n}\sum_{s=1}^{n}(y_s - \hat{y}_{s,cv})^2$. We select the model

for which this quantity is smaller. The second method is significance testing. From the data, we test whether β is significantly different from 0. If it is, then we use model (M2) for prediction, if not, we use (M1).
For each of the 10,000 simulated data sets, we performed both methods to decide between the two models. We chose $x^* = 0.3$ as the point at which we wanted to predict. Then we calculated the predictions from both models, and compared how well the prediction fitted to $10+x^*$, the conditional expectation of y at x^*.
We found that among the 10,000 data sets, in all cases β was significant, while cross-validation correctly decided for the regression-model (M2) only in 3,497 of the 10,000 experiments. So there were 3,497 cases where significance testing and cross-validation decided for the same model. In the remaining 6,503 cases, there were 213 when the simpler model (M1) chosen by cross validation gave better prediction, while there were 6,290 cases when model (M2) chosen by significance testing gave better predictions in x^*.
Things were even more extreme when for the same 10,000 simulated data sets we chose to predict at $x^* = 5$. Then the prediction from the regression model (M2) was better in all 10,000 cases. Therefore, significance testing chose the better model in 6,503 cases, while cross-validation never chose a model which led to a better prediction than the one chosen by significance testing.

The results in Example 1 need some comments. It is clear that the poor performance of the cross validation is due to the fact that there is just one single x_i which is far away from the others. With an observational study, we would usually not have just one observation for which the x-value is far away from the others. If we had, then we might decide that observation 7 is an outlier and not use this observation for prediction at all. Therefore, the prediction from model (M1) would fit much better to observations 1 to 6 (and give much better prediction for $x^*=0.3$). With an observational study which observes x_i's in the range between 0 and 1, we would usually not want to make predictions for $x^* = 5$.

The example seems to show that cross validation is not appropriate for designed experiments. This, however, is only part of the story. Had the experiment been properly designed then cross validation would not produce such results.

3.3 Optimal conditions for **c**ross validation

To get an impression of how cross validation performs with well-designed experiments, we will look for optimal conditions for cross validation, and check for what designs such conditions are met.

Assume that we have a matrix $X = [x_1, x_2, ..., x_n]^T$, that is x_i^T is the ith row of X, and we have a vector $x^*=[x^*_1,...,x^*_n]^T$ such that for the target vector y of observations from our experiment it holds

$$y = X\beta + x^* \gamma + \varepsilon$$

where β is a (K+1)-dimensional vector and γ a constant, while ε is a random vector, such that E(ε) = 0 and Cov(ε) = $\sigma^2 I_n$. We consider RSS_{cv} when the model

$$y = X\beta + \varepsilon$$

is assumed, i.e. when the factor associated with γ is neglected in the model. Thus,

$\hat{y} = X(X^T X)^{-1} X^T y$ is the least squares estimate for y from the assumed model.

We start with some definitions. Let $\omega^{\perp}(X) = I_n - X(X^T X)^{-1} X^T$, $d_i = 1 - x_i^T (X^T X)^{-1} x_i$ the i^{th} diagonal element of $\omega^{\perp}(X)$, and $D = \text{diag}(d_1,...,d_n)$. Let $\hat{y}_{cv} = [\hat{y}_{1,cv},..., \hat{y}_{n,cv}]^T$ the vector of predictions of y from leave one out cross validation.

With these definitions we have (see e.g. Cook and Weisberg, 1982, p. 33) that

$$y - \hat{y}_{cv} = D^{-1}(y - \hat{y}) .$$

As $y - \hat{y} = \omega^{\perp}(X) y$ and as $y = X\beta + x^* \gamma + \varepsilon$, it follows that

$$y - \hat{y}_{cv} = D^{-1}\omega^{\perp}(X) x^* \gamma + D^{-1}\omega^{\perp}(X)\varepsilon .$$

Therefore,

$$\text{E}(RSS_{cv}) = \text{E}(y - \hat{y}_{cv})^T (y - \hat{y}_{cv}) = \text{E}((\varepsilon + x^* \gamma)^T \omega^{\perp}(X) D^{-2} \omega^{\perp}(X)(\varepsilon + x^* \gamma)) .$$

Due to the fact that E(ε) = 0, we get

$$\text{E}(RSS_{cv}) = \text{E}(\varepsilon^T \omega^{\perp}(X) D^{-2} \omega^{\perp}(X)\varepsilon + \gamma^2 x^{*T} \omega^{\perp}(X) D^{-2} \omega^{\perp}(X) x^*) . \qquad (1)$$

The first term in (1) can be transformed to

$$\text{E}(\varepsilon^T \omega^{\perp}(X) D^{-2} \omega^{\perp}(X)\varepsilon) = \text{tr}\left(\omega^{\perp}(X) D^{-2} \omega^{\perp}(X)\, \text{E}(\varepsilon\varepsilon^T)\right)$$

$$= \text{tr}\left(\omega^{\perp}(X) D^{-2} \omega^{\perp}(X)\sigma^2 I_n\right) = \text{tr}\left(D^{-2}\omega^{\perp}(X)\right)\sigma^2 = \sigma^2 \sum_{i=1}^{n} \frac{1}{d_i},$$

because d_i is the i^{th} diagonal element of $\omega^{\perp}(X)$ and because D^{-2} = diag($1/d_1^2,...,1/d_n^2$).

In the situation that $\gamma = 0$, i.e. when the fitted model is appropriate, then it is desirable that E(RSS_{cv}) is as small as possible. Using that $\sum d_i = \text{tr}\ \omega^{\perp}(X) = n - K - 1$, we get

$$\sum_{i=1}^{n} 1/d_i \geq \sum_{i=1}^{n} \frac{n}{n-K-1} = \frac{n^2}{n-K-1}$$

with equality holding if all d_i are equal. Hence, in the special instance that $\gamma = 0$ and that all d_i are equal,

$$\text{E}(RSS_{cv}) = \sigma^2 \frac{n^2}{n-K-1} \quad \text{is minimum.}$$

In this case, we can see an advantage of RSS_{cv} compared to RSS, the usual sum of squares for errors: since n^2 / (n - K -1) is increasing in K, we have E(RSS_{cv}) increasing if some additional irrelevant parameters are added to the fitted model.

In the case $\gamma \neq 0$, we consider the second term in (1). Defining $[a_1,...,a_n] = x^{*T}\,\omega^{\perp}(X)$, we get

$$\gamma^2 x^{*T}\,\omega^{\perp}(X) D^{-2} \omega^{\perp}(X) x^* = \gamma^2 \sum_{i=1}^{n} a_i^2 / d_i^2 .$$

Note that $\sum a_i^2 = x^{*T}\,\omega^{\perp}(X) x^* \leq x^{*T}\,x^*$, with equality if $x^{*T}X = 0$. In the case that $\gamma \neq 0$, we want E(RSS_{cv}) to be as large as possible. If all d_i are equal, we can achieve this for $x^{*T}X = 0$, i.e. if x^* is orthogonal to the factors in the fitted model.

So, if all d_i are equal, it is true that

$$\mathrm{E}(RSS_{cv}) = \sigma^2 \frac{n^2}{n-K-1} + \gamma^2 \frac{n^2}{(n-K-1)^2} x^{*T}\,\omega^{\perp}(X) x^* \leq \sigma^2 \frac{n^2}{n-K-1} + \gamma^2 \frac{n^2}{(n-K-1)^2} x^{*T}\,x^*$$

with equality if x^* is orthogonal to the factors in the fitted model.

If we compare this to the corresponding formula for the case that x^* is included in the fitted model

$$\mathrm{E}(\tilde{RSS}_{cv}) = \sigma^2 \frac{n^2}{n-K-2},$$

then we see that $\mathrm{E}(\tilde{RSS}_{cv}) \leq \mathrm{E}(RSS_{cv})$ if

$$\sigma^2 \frac{n^2}{n-K-2} \leq \sigma^2 \frac{n^2}{n-K-1} + \gamma^2 \frac{n^2}{(n-K-1)^2} x^{*T}\,x^* .$$

Some algebra shows that this is equivalent to

$$\sigma^2 \frac{1}{n-K-2} \leq \gamma^2 \frac{1}{n-K-1} x^{*T}\,x^* .$$

So the performance of the refined model with x^* included in comparison to the simpler model depends on the size of $\gamma^2 x^{*T}\,x^*$ compared to σ^2.

However, it should be noted that this holds only if

a) for both models all d_i are equal, and
b) x^* is orthogonal to the columns of X, i.e. to the factors contained in the simpler model.

Note that condition a) formulates a sort of balance in the observations since it demands that all observations x_i^{T} have the same norm corresponding to some Mahalanobis distance.

The poor performance of cross validation in Example 1 can be explained easily now: in the more elaborated model the d_i were highly dissimilar. In fact, $d_7 = 0.002$, while $d_i = 0.2$ for the other observations. In contrast, all d_i would have been equal for the linear regression model, if half of the x_i had been $+a$ or $-a$ each, for some appropriate number a. The second term of (1) then is maximised if a is as large as possible (for constant effect γ). Interestingly, this is the D-optimal linear regression design (see, e.g. Pukelsheim, 1993, p. 57). More designs with the properties a) and b) are given in section 4.

3.4 Variables selection by means of cross validation

For **variables selection** properties a) and b) have a very important impact. They imply that the second term of formula (1) is proportional to $\gamma^2 \cdot x^{*T} x^*$, where γ is the effect of the factor not included in the fitted model. Therefore, the size of RSS_{cv} is reduced the most if the factor with the largest $|\gamma| \sqrt{x^{*T} x^*}$ is included in the model. If then, as in famous screening designs (see section 4), all $x^{*T} x^*$ are equal, this is the factor with largest $|\gamma|$. Therefore, for such designs greedy variables selection can be based on the absolute value of the estimates of the unknown coefficients corresponding to the possibly influential factors. Hence, it is not necessary to determine R^2_{cv} for all possible factors individually to identify the next factor to be included in the model. Instead, this factor can be fixed just by inspection of the coefficients determined by estimating the model where all factors are included. R^2_{cv} is only used as a stop criterion, i.e. to determine the model size.

Unfortunately, inspection of this **stopping rule** shows that conditions a) and b) might also imply that greedy variables selection is always choosing the maximum model if the number of designed factors is maximum. This can be seen as follows.

Let us assume for the moment that the number of already chosen factors K is so small that $n > K+2$, then from condition a) it follows for a model with K factors that

$$y - \hat{y}_{cv} = D^{-1}(y - \hat{y}) = \frac{n}{n-K-1}(y - \hat{y}).$$

Therefore,

$$RSS_{K,cv} = \frac{n}{n-K-1} RSS_K ,$$

where $RSS_{K,cv}$ and RSS_K stand for the residual sum of squares after K factors have been included in the model.

The stopping rule would stop variables selection iff cross validated R^2 is smaller after the $(K+1)^{th}$ step than after the K^{th} step, i.e. iff inclusion of one more variable would decrease performance, i.e.

$$R^2_{K+1,cv} < R^2_{K,cv} ,$$

which is equivalent to

$$RSS_{K+1,cv} > RSS_{K,cv} .$$

In the case of all d_i equal the above relationship of RSS_{cv} and RSS, thus, leads to

$$RSS_{K+1} > \frac{n\text{-}K\text{-}2}{n\text{-}K\text{-}1} RSS_K .$$

Let us now assume a situation where all factors in the design are orthogonal (extending condition b)). If $SSC_{K+1} := \gamma^2_{K+1} SS_{K+1}$ is defined as the contribution of factor $(K+1)$ to the

overall sum of squares of the target, then for orthogonal factors these contributions are summing up to the overall sum of squares, and thus

$$RSS_{K+1} = RSS_K - SSC_{K+1} > \frac{\text{n - K - 2}}{\text{n - K - 1}} RSS_K$$

which is obviously equivalent to

$$RSS_K > (n-K-1)SSC_{K+1} \ . \qquad (2)$$

This leads, however, to a contradiction if the number of designed factors is equal to *n-1*. In such a design there are exactly (*n-K-1*) factors remaining as candidates for inclusion in the model with contributions at most that high as for the factor to be chosen in step (*K+1*) since this is the factor with minimum RSS_{cv} and thus *RSS* of the factors remaining for selection. Moreover, ultimately *RSS* will be zero because of the orthogonality of the factors, and therefore RSS_K has to be equal to the sum of the contributions of all remaining (*n-K-1*) factors. This is obviously not possible if inequality (2) is valid.

The only case remaining to be discussed is when we have already $K = n-2$, i.e. $K+1 = n-1$ and thus only one factor is remaining. In this case, RSS_{K+1} and thus $RSS_{K+1,cv}$ has to be zero, and thus R^2_{cv} takes its maximum value, and also this factor is chosen, if it generates a nonzero contribution. Thus, if the number of designed factors is maximum for designs where condition a) is valid and all factors are orthogonal, the stopping rule is only eliminating factors with zero contribution.

On the other hand, the stopping rule potentially works if the number of designed factors is not maximum, i.e. not equal to *n-1*. E.g., imagine that only *n-2* factors are designed and that the contributions of the remaining *n-K-2* factors not included yet in the model are equal and smaller than the contribution of the error term of the linear model with all designed factors. Then inequality (2) is obviously fulfilled, and none of the remaining factors will be included in the model. If the contribution of the factor proposed in the $(K+1)^{th}$ step is bigger than the contributions of the other remaining factors, the error term has to be even bigger in order to stop factor inclusion in the model, i.e. variables selection.

4. Greedy variables selection in experimental design

In this section examples of variables selection in experimental studies are discussed. In particular, we will examine screening and optimisation experiments, and discuss their suitability for variables selection by means of cross validation.

4.1 Screening

In screening, **linear models** are used with coded influential factors (see Weihs and Jessenberger, 1999).

Screening model and designs

A **screening model** is of the form:

$$y_i = \beta_1 + \sum_{j=1}^{K} x_{c_{ij}} \beta_{j+1} + \varepsilon_i, \varepsilon_i \sim i.i.N(0, \sigma^2),$$

where y_i is the result of the target y in the i^{th} trial, $x_{c_{ij}}$ the coded level of the j^{th} factor in the i^{th} trial, β_1 the increment, β_{j+1} the half effect of the j^{th} factor on y, ε_i the error in the i^{th} trial and σ^2 the error variance. In matrix form one can write:

$$y = X\beta + \varepsilon, \text{ where } X = \begin{pmatrix} 1 & \\ \vdots & A \\ 1 & \end{pmatrix}$$

is the **design matrix** including the **plan matrix** A with the column representation $A = (x_{c_1} \ldots x_{c_K})$, where $x_{c_j} = (x_{c_{1j}} \ldots x_{c_{nj}})^{\mathrm{T}}$.

I.e. X is a matrix with all ones in the 1st column and the columns of the matrix A.

$\beta := (\beta_1\ \beta_2 \ldots \beta_{K+1})^{\mathrm{T}}$ is the vector of the unknown **model coefficients** and

$\varepsilon := (\varepsilon_1 \ldots \varepsilon_n)^{\mathrm{T}}$ is the error vector.

The n rows of the plan matrix A correspond to the n trials, the K columns to the controlled factors. Each factor takes only two levels which are coded −1 and +1.

Such a plan matrix defines a **screening plan** iff the coded factors all have mean 0 and are pairwise orthogonal, i.e. $x_{c_j}^T x_{c_k} = 0, j \neq k$.

Note that in a screening plan A each column consists of exactly as many −1 as +1 in order to guarantee mean 0 for each factor, and that from these two properties it follows that $X^T X = n \cdot I$. Note that such screening plans are D-optimal (see, e.g. Cheng, 1980).

The aim of screening is **factor reduction**. Thus, the question is how to distinguish between relevant and irrelevant factors. This is a natural task for a variable selection method. From the definition of screening designs it can be easily seen that all such designs fulfil the conditions a) and b) in section 3.3 so that they are very well suited for leave one out cross validation. Moreover, the shortcut version of greedy variable selection based on the size of factor effects can be used as motivated in section 3.4.

Greedy variable selection in screening experiments

In screening experiments, **stepwise regression with greedy forward selection** first selects the factor with the largest absolute value of the regression coefficient, then the factor with second largest absolute value of the regression coefficient, etc. until the predictive power is no more increasing.

Famous examples for screening designs for which conditions a) and b) in section 3.3 are fulfilled are the fractional factorial designs (see e.g. Weihs and Jessenberger, 1999). With a fractional factorial design for every submodel with $K \leq n$-2 factors (or even interactions, not discussed here) we have $X^TX = n\ I_{K+1}$ (cp. Table 1 for a coded fractional factorial design for 4 factors generated from a full factorial with 16 runs to have the property ABCD=1). It follows that $x_i^T (X^T X)^{-1} x_i = (K+1)/n$, independent of i, because each x_i consists of K+1 elements which are either +1 or -1. Additionally, we have for each additional factor or interaction that x^* is orthogonal to every column of X. Another class of screening designs with the same properties are Plackett-Burman designs (see e.g. Weihs and Jessenberger, 1999). An example of a coded Plackett-Burman design with 8 runs for 7 factors A – G can be found in Table 2. This design (except the 1st line and column) is generated from the line marked in bold by cyclical right shift.

Table 1: Fractional factorial design

Const.	A	B	C	D
1	−1	−1	−1	−1
1	+1	+1	−1	−1
1	+1	−1	+1	−1
1	−1	+1	+1	−1
1	+1	−1	−1	+1
1	−1	+1	−1	+1
1	−1	−1	+1	+1
1	+1	+1	+1	+1

Table 2: Plackett-Burman design

Const.	A	B	C	D	E	F	G
1	−1	−1	−1	−1	−1	−1	−1
1	**−1**	**−1**	**+1**	**−1**	**+1**	**+1**	**+1**
1	+1	−1	−1	+1	−1	+1	+1
1	+1	+1	−1	−1	+1	−1	+1
1	+1	+1	+1	−1	−1	+1	−1
1	−1	+1	+1	+1	−1	−1	+1
1	+1	−1	+1	+1	+1	−1	−1
1	−1	+1	−1	+1	+1	+1	−1

Unfortunately, inspection of the **stopping rule** in section 3.4 shows that for saturated orthogonal designs like the standard version of Plackett-Burman designs, for which the number of designed factors is equal to *n-1*, greedy variables selection is always choosing

the maximum model. This implies that Plackett-Burman and fractional factorial designs are only optimal with respect to cross validation if the number of involved factors is not maximum, i.e. if there are degrees of freedom left also when all designed factors are included in the model. This is true for the fractional factorial design in Table 1. For the Plackett-Burman design in Table 2 this can be achieved, e.g., by eliminating at least one of the columns A – G.

4.2 Optimisation

Near to an optimum, e.g. inside an 'inverted cup' region, the target cannot be represented adequately by a linear model in the influential factors alone, one needs a **quadratic model** in the factors, at least. As an optimisation model we, thus, use a quadratic model in those factors which were selected to be relevant for the target in earlier stages of experimentation (see Weihs and Jessenberger, 1999). Note that in what follows we restrict attention to quantitative influential factors.

Optimisation model and designs

In an **optimisation model**, the target is modelled as a linear model in the coded factors, in their two-factor interactions, and in their squares:

$$y_i = \mu + \sum_{j=1}^{K} x_{cij}\,\beta_j + \sum_{j=1}^{K-1}\sum_{k>j} x_{cij}x_{cik}\,\beta_{j,k} + \sum_{j=1}^{K} x_{cij}^2\,\beta_{j,j} + \varepsilon_i,\ \varepsilon_i \sim i.i.N\left(0,\sigma^2\right),$$

where y_i is the result of the target in the i^{th} trial, x_{cij} the coded level of the j^{th} factor in the i^{th} trial, μ the intercept (overall mean), β_j the coefficient of the j^{th} factor, $\beta_{j,k}$ the coefficient of the interaction of the j^{th} with the k^{th} factor, $\beta_{k,k}$ the coefficient of the squared k^{th} factor, ε_i the error in the i^{th} trial and σ^2 the error variance.

A plan matrix defines an **optimisation plan** iff all the involved coded factors take at least three different levels.

Note that for optimisation models it is neither assumed that the coded factors only take values -1 and $+1$ and have mean 0, nor that for the design matrix X it is true that $X'X = n \cdot I$. Therefore, the least squares estimates of the model coefficients are not interpretable as (half) effects of the factors, interactions or squared factors. The only condition a plan matrix has to fulfil is that all the involved factors take at least three different levels, because otherwise the effect of the squared factors is not estimable.

For the selection of optimisation plans it is particularly important that the target is observed in sufficiently many points in the region of interest in order to be able to estimate the coefficients of the quadratic model reliably. Since the optimum will probably lie in a point of the region of interest in which no trial was carried out, the model has to be valid in the whole region. In order to avoid overfitting, we propose variables selection also for

optimisation models of the above kind. Since the aim of optimisation modelling is a good prediction of the target in the optimum, at least, we propose to use predictive power, i.e. R^2_{CV}, as the selection criterion.

The results of section 3.3 indicate that for variables selection it is important to choose the experimental design in such a way that for each model to be compared the d_i are as equal as possible. Unfortunately, in optimisation neither this condition a) nor the orthogonality of the factors in condition b) will be fulfilled in general.

We will now compare the performance of two standard optimisation designs with respect to variable selection by means of cross validation. Let us begin with an example of a design with a particularly poor performance. Assume we have three factors, and we want to run an efficient design in 15 runs. In this case, a standard design is a rotatable central composite design where the factors are set as in Table 3a (cp. Box and Draper, 1987, 512). Note that the design entries are coded in that they are standardised to lie between –1 and 1. An interesting feature is that for this design $d_9 = 0$ for all models that contain all three quadratic effects (cp. Table 4). Therefore, the model choice with the help of cross validation does not work properly for this design: No matter how big the effects of the squared factors may be, cross validation will never select a model with all three of them. This would be different, if the run in the centre point (run 9) is doubled, though.

Table 3: Optimisation designs

	a) rotatable			b) inscribed		
Run	A	B	C	A	B	C
1	1	1	1	1	1	1
2	1	1	-1	1	1	-1
3	1	-1	1	1	-1	1
4	1	-1	-1	1	-1	-1
5	-1	1	1	-1	1	1
6	-1	1	-1	-1	1	-1
7	-1	-1	1	-1	-1	1
8	-1	-1	-1	-1	-1	-1
9	0	0	0	0	0	0
10	1.68	0	0	0.85	0	0
11	-1.68	0	0	-0.85	0	0
12	0	1.68	0	0	0.85	0
13	0	-1.68	0	0	-0.85	0
14	0	0	1.68	0	0	0.85
15	0	0	-1.68	0	0	-0.85

In the second design with 15 runs the d_i are not that different (cp. Table 4) and therefore the model search with RSS_{cv} works reasonably well. This so-called inscribed central composite design was chosen in order to be able to use results from an earlier screening experiment so that the augmented design does not exceed the original experimental region. Table 3b shows such a design the first eight trials of which build a full factorial screening design.

Table 4: Expected RSS_{cv} if there are all active effects included in the respective model

Factors in the model	Bound	Inscribed design	Rotatable design
A	17.31	17.36	17.42
A, B	18.75	18.97	18.90
A, B, C	20.45	21.07	20.54
A, B, C, AB	22.5	24.38	22.88
A, B, C, A²	22.5	23.07	23.57
A, B, C, AB, AC	25	29.95	26.47
A, B, C, A², B²	25	25.26	26.89
A, B, C, AB, AC, BC	28.12	41.31	32.67
A, B, C, A², B², C²	28.12	28.65	∞
A, B, C, AB, AC, BC, A²	32.14	49.80	35.71
A, B, C, AB, AC, BC, A², B²	37.5	53.72	39.08
A, B, C, AB, AC, BC, A², B², C²	45	57.91	∞

Table 4 gives an overview over the performance of cross validation for different model choices when either the inscribed design or the rotatable design is used. The entries in Table 4 are the $\sum 1/d_i$, which is proportional to the first term in formula (1) in section 3.3. Hence, they indicate the expected size of RSS_{cv} if all the relevant factors are already in the model. For comparison we also give the lower bound $n^2/(n-K-1)$ which is achieved by an ideal design with all d_i equal. Note that if two models differ too much in $\sum 1/d_i$, then we will decide for the model with the smaller $\sum 1/d_i$, even if the other model produces a much better fit.

Table 4 shows that for all models the inscribed design promises to provide a reasonable RSS_{cv}, while the rotatable design with 15 runs collapses for the models which contain all quadratic effects. Note that the inscribed design prefers models with fewer interactions if the number of factors is fixed.

5. Conclusion

In this paper we discussed the pros and cons of cross validation for variables selection using experimental design. On the one hand, we illustrated that experimental studies provide a better basis for cross validation than observational studies, since the properties of the observed factors can be controlled. Design properties favorable to greedy variables selection are identified, namely a certain balance in the observations and orthogonality of the factors. Screening designs meet these properties. On the other hand, however, for special screening designs, namely saturated orthogonal designs, it was shown that cross validation does a very poor job on variables selection since it only eliminates variables with no contribution to the target at all. Finally, it was demonstrated that otherwise optimal designs like rotatable designs may be sub-optimal for cross validation.

Acknowledgement

This work has been supported by the Collaborative Research Centre "Reduction of Complexity in Multivariate Data Structures" (SFB 475) of the German Research Foundation (DFG).

References

Box, G.E.P., and Draper, N.R. (1987): *Empirical Model Building and Response Surface* , New York: Wiley

Cheng, C.S. (1980): Optimality of Some Weighting and 2^n Fractional Factorial Designs. *Annals of Statistics* **8**, 436 – 446.

Cook, R.D. and Weisberg, S. (1982): *Residuals and Influence in Regression*, London: Chapman and Hall.

Gelfand, A.E. (1996): Model Determination Using Sampling-based Methods; in: Gilks, W.R., Richardson, S., and Spiegelhalter, D.J: Markov Chain Monte Carlo in Practice; Chapman & Hall, London, 145-161

Pukelsheim, F. (1993): *Optimal Design of Experiments*, New York: Wiley.

SAS/STAT User's Guide (1990): Version 6, Fourth Edition, Volume 2, The SAS Institute, Cary, NC.

Weihs, C. (1993): Multivariate Exploratory Data Analysis and Graphics: A tutorial, *Journal of Chemometrics* **7**, 305-340

Weihs, C., Jessenberger, J. (1999): *Statistische Methoden zur Qualitätssicherung und -optimierung in der Industrie*. Weinheim: Wiley-VCH.

Weiss, S.M. and Kulikowski, C.A. (1991): Computer Systems that Learn; Morgan Kaufmann, San Francisco

Weiss, S.M. and Indurkhya, N. (1998): Predictive Data Mining; Morgan Kaufmann, San Francisco