A novel approach for screening discrete variations in organic synthesis

Rolf Carlson^{1*}, Johan Carlson² and Anders Grennberg³

¹*Department of Chemistry, University of Tromsø, N-9037 Tromsø, Norway*

²*Division of Industrial Electronics, Luleå University of Technology, SE-971 87 Luleå, Sweden*

³*Division of Signal Processing, Luleå University of Technology, SE-971 87 Luleå, Sweden*

SUMMARY

In this paper we present a general strategy for screening discrete variations in organic synthesis. The strategy is based upon principal properties, i.e. principal component characterization of the constituents defining the reaction system. The first step is to select subsets of test items from each class of constituents defining the reaction space, i.e. substrates, reagents, solvents, catalysts, etc., so that the selected items from each class cover the properties considered. The second step is to construct a candidate matrix which contains all possible combinations of the items in the subsets. This matrix is a full multilevel factorial design. The third step is to assign a tentative model for the screening experiment and to construct the corresponding candidate model matrix. The fourth step is to select experiments to yield an experimental design that spans the variable space efficiently and that also gives good estimates of the model parameters. We present an algorithm that uses singular value decomposition to select experiments. The proposed strategy is then illustrated with an example of the Fischer indole synthesis. Copyright © 2001 John Wiley & Sons, Ltd.

KEY WORDS: optimal experimental design; principal properties; organic synthesis; Fischer indole synthesis; discrete variations; combinatorial chemistry

1. INTRODUCTION

Organic synthesis is an experimental science. An important field of research is the development of efficient procedures for carrying out specific transformations. This often involves tedious work to elaborate reactions into efficient and optimum synthetic methods. There are two general problems encountered in this process. First, it will be necessary to determine which combination of reagent(s), co-reagent(s), catalyst, solvent, etc. will afford the most promising result and thus merit further exploration. Second, how should the detailed experimental conditions in this reaction system be adjusted to obtain an optimum result? For the second type of problem there is an abundance of established techniques for designing and analyzing experiments. For the first type of problem, strategies are still in their infancy. The present paper addresses this problem.

* Correspondence to: Rolf Carlson, Department of Chemistry, University of Tromsø, N-9037 Tromsø, Norway.

E-mail: rolf@chem.uit.no

Contract/grant sponsor: Norwegian Research Council

Figure 1. Essential features of the problem, described by the reaction space.

We present a sequential strategy for the systematic exploration of discrete variations in organic synthesis.

2. PROBLEM DESCRIPTION

The essential features of the problem can be described using the *reaction space* depicted in Figure 1. The *axes* of the reaction space define *variations* in the nature of the different constituents (substrate, reagents, solvent, etc.) that define the reaction system. The entire reaction space is defined as the union of all possible combinations of the constituents. An experimental design for exploring the reaction space defines a selection of test systems that cover the space as efficiently as possible. In a screening experiment, where the objective is to find suitable design systems for further elaboration, a design which gives a large spread of design points in the reaction space is appropriate. If the objective is to analyze how a gradual change in the properties of the constituents in the reaction system exerts an influence on the result, a design with a more uniform distribution of design points would be a suitable choice.

The number of experiments to be run will of course depend on the complexity of the question posed to the experimental system. If the objective is to find a combination of reagent, co-reagent, catalyst and solvent that can carry out a specific transformation, e.g. a critical step in a multistep synthesis, the goal is reached when a useful combination has been found. For this reason a good design strategy must be sequential and interruptible after each experimental run. If the objective is to determine the scope and limitations of a new reaction, the design must select test items so that the critical properties of the system can be identified. This will require a series of experiments to ensure that the systematic variation in the system has been explored.

To tailor a detailed strategy for this type of design, we need to quantify the *axes* of the reaction space.

2.1. Principal properties

This can be accomplished through the *principal properties* for the classes of the various constituents. A thorough discussion of this concept is given in the book by Carlson [1]. The following is only a brief outline.

When a molecule takes part in a reaction or when it interacts with its surroundings, the outcome is determined by the properties at the molecular level. However, such properties cannot be measured directly. Some molecular properties can be estimated through quantum chemical modeling, while others can be linked to observed properties through physical chemical models. Generally speaking, molecules can be characterized by their properties, and observable properties can likely be assumed

to be observable manifestations of intrinsic molecular properties. In several respects the members of the classes of constituents defining the *axis* in reaction space can be assumed to be similar. For example, if a reaction is to be developed with ketones as substrates, the class of ketones defines the set of substrates. They share the same functional group and are in this respect similar. The variations in the nature of the side chains of the ketones make the various ketones slightly different with respect to their observable properties. Assuming that the observable properties can be used as probes of the intrinsic properties, we can analyze the variation in the intrinsic properties. We can also assume that observable properties which depend on the *same* intrinsic property will be more or less correlated with each other over the set of molecules and that observable properties which depend on *different* intrinsic properties will not be strongly correlated. An analysis that takes these assumptions into account can be carried out through principal component modeling of the observed variation in the properties over a set of ketones. Assume that each ketone is characterized by a set of k molecular property descriptors, $\{q_1, q_2, \dots, q_k\}$. Assume also that these descriptors have been chosen so that they are likely to pick up those molecular properties we believe will exert an influence on the outcome of the reaction to be studied. We shall of course use all background knowledge and all prior information to select relevant descriptors. This means that *which* descriptors to choose will depend on *which* reaction we intend to study.

A set of n ketones will define an $n \times k$ descriptor matrix $\mathbf{Q}$. For convenience, assume also that $\mathbf{Q}$ has been mean centered and that each descriptor has been scaled to unit variance. This matrix displays two types of variation: vertically, it shows the variation in the properties between ketones; horizontally, it shows the variation over a set of ketones between descriptors. These variations can be isolated and displayed by a principal component factorization of $\mathbf{Q}$. This yields

$$\mathbf{Q} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \quad (1)$$

where $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_A]$ is the score matrix defined by the eigenvectors $\mathbf{t}_i$ ($i = 1, \dots, A$) of the correlation matrix $\mathbf{Q}\mathbf{Q}^T$; the matrix $\mathbf{P}^T$ is the transpose of the loading matrix defined by the eigenvectors $\mathbf{p}_i$ ($i = 1, \dots, A$) of the variance-covariance matrix $\mathbf{Q}^T\mathbf{Q}$ ($\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_A]$). A matrix of residuals, $\mathbf{E}$, will occur when the number of components $\mathbf{p}_i$ included in the model is less than the number of original descriptor variables.

The number of significant components, A , can be determined by cross-validation [2]. The principal component analysis constitutes a projection of the n object points in the k -dimensional descriptor space onto an A -dimensional hyperplane spanned by the eigenvectors $\mathbf{p}_i$. The elements t_{ij} of the score vectors $\mathbf{t}_i$ are orthogonal projections of the object j on the eigenvector $\mathbf{p}_i$. The principal component model describes the systematic variation in the properties over a set of compounds. The non-systematic random variation is collected in the residual matrix $\mathbf{E}$. The benefit of this procedure is that the number of variables we need to consider for taking the systematic variation into account will be less than the number of original descriptor variables. We use the term *principal properties* to define the correlative pattern of the original descriptors as portrayed by the eigenvectors of property matrices. As the eigenvectors are orthogonal, we can assume that they reflect independent intrinsic properties. We can then use the eigenvectors to define the *axes* of the reaction space and the corresponding score values t_{ij} to quantify the variation along these *axes*. A plot of the score vectors against each other, a score plot, e.g. $\mathbf{t}_1$ vs $\mathbf{t}_2$, displays the scatter of the object points projected onto the first two eigenvectors. Objects that are close to each other in the k -dimensional descriptor space will be close to each other in the score plot. Conversely, objects that are dissimilar to each other and hence located far from each other in the descriptor space will be projected far from each other in the score plot. This forms the basis of the present strategy. When there are only a few *dimensions* of the reaction space to consider, a selection of the corresponding test candidates can be made from a visual

inspection of the corresponding score plots. A discussion is given by Carlson [1]. In the general case, however, the complexity of the problem increases rapidly as the number of *dimensions* in the reaction space increases. To obtain a suitable design for the selection of test items in a screening experiment, the design should fulfil certain desired criteria. In the next section we outline such a strategy.

2.2. Scope of the problem

Assume that we wish to explore a condensation reaction between a ketone {*substrate*} and an active methylene compound $\text{CH}_2\text{Z}^1\text{Z}^2$, where Z^1 and Z^2 are electron-withdrawing groups {*reagent*} in the presence of a Lewis acid {*catalyst*} and a proton-abstracting Brønsted base {*co-reagent*} in a solvent. The reaction space is spanned by {*substrate*}, {*reagent*}, {*co-reagent*}, {*catalyst*} and {*solvent*}. If we limit the selection to commercial available chemicals, we can choose among approximately 80 ketones, 30 active methylene compounds, 120 Lewis acid catalysts, 60 Brønsted bases and 150 common solvents. This yields roughly 2.6×10^9 possible combinations, a prohibitively large number, and therefore an exhaustive study is precluded.

Assume that each class of constituents has been characterized by its principal properties determined from descriptors *we consider to be relevant for the problem under study*. For convenience we can assume that two principal components are sufficient for each class. From the corresponding score plots it is then possible to select a subset of test items from each constituent class so that the selected test items span the range of variation of interest. With two principal components this is easily done by visual inspection. Assume that 10 items from each class have been chosen. This yields a total of 1.0×10^5 possible combinations, which is still a prohibitively large number. Often the situation is less complicated than in the above examples. A common situation is that the experimenter wishes to analyze, for example, combinations of solvents and catalysts for a series of test substrates. Nevertheless, even fairly simple cases can easily yield a large number of possible combinations, often in the range of thousands to some tens of thousands of possible combinations, which of course also precludes an exhaustive study.

We suggest that these difficulties can be overcome by using the principal property score values of the items in the selected subsets of possible test items as design variables and constructing a design with desired properties from these variables. For this we need a model that relates the outcome of the experiment to the properties of the experimental system.

2.3. Models for screening

Let y be the value of the measured response, i.e. the outcome of the experiment, e.g. yield, a measure of selectivity or the specific rate. It is reasonable to assume that the value of y will depend on the nature of the reaction system and that there is some kind of functional relationship between y and the properties of the system, i.e.

$$y = f(\text{properties}) + \varepsilon \quad (2)$$

Since the principal property scores are measures of the molecular properties we assume to be relevant, we can assume that the essential variation in y can be modeled as a function of the principal properties, i.e.

$$y = f(\text{principal properties}) + \varepsilon \quad (3)$$

Assume that k principal properties are sufficient to span the systematic variation in the reaction space and let x_i ($i = 1, \dots, k$) be the corresponding score values. The functional relationship can then be

expressed as

$$y = f(x_1, x_2, \dots, x_k) + \varepsilon \quad (4)$$

It is difficult, however, to derive an analytical expression for f based upon purely theoretical considerations, but an approximate model can be determined by the following reasoning.

The outcome of a chemical reaction depends on the energetics of the chemical system. The transformation of starting materials into reaction products can be described as a movement on a potential energy surface, from one minimum defined by the starting materials to another minimum defined by the products. The energy barrier to surmount is related to the rate of the reaction, and the energy differences between the energy minima are related to equilibria. Rates and equilibria determine the results of interest in synthetic chemistry. How shallow the minima are and how high the energy barriers are will depend on the properties of the constituents of the reaction system and the detailed experimental conditions. Perturbation of the properties of the reaction system will change the positions of the minima and maxima of the potential energy surface and hence alter the outcome of the reaction. It is reasonable to assume that a gradual change in the properties will produce a gradual change in the shape of the potential energy surface. For this reason we assume that for screening experiments a sufficiently good approximation of f can be accomplished by a truncated Taylor expansion of f expressed in the principal property variables x_i :

$$f = \beta_0 + \sum_{i=1}^k \beta_i x_i + \sum_{i=1}^k \sum_{j=1}^k \beta_{ij} x_i x_j + \sum_{i=1}^k \sum_{j=1}^k \sum_{l=1}^k \beta_{ijl} x_i x_j x_l + \text{higher-order terms} \quad (5)$$

How many terms to include in the model will depend on how detailed information is desired. A model with only linear terms $\beta_i x_i$ makes it possible to determine whether or not variations in the properties of the constituents have any significant influence at all and which properties are important. If the model is augmented with cross-product terms $\beta_{ij} x_i x_j$, it is possible to evaluate interaction effects between constituents of the reaction system. Such interactions are always to be expected. An example will illustrate this. If the solvent is changed, the solvation of dissolved species is changed and this is likely to influence the reactivity and hence the outcome of the reaction. Since the extent of solvation and the nature of such interactions between the solute and the solvent are dependent on the properties of the solvent and the properties of the solute, we can expect interactions between the corresponding principal properties. This can be accounted for mathematically by a corresponding cross-product term in the model. If also square terms $\beta_{ii} x_i^2$ are included, it is possible to describe non-linear effects. We strongly recommend that interaction effects should always be considered when scope and limitations are to be determined.

The roles played by the principal properties of the constituents of the reaction space can be assessed from the values of the model parameters β_0 , β_i , β_{ij} and β_{ii} . The problem is thereby reduced to finding an experimental design from which the model parameters can be estimated. We assume that in the general case a second-order Taylor expansion will be sufficient, i.e.

$$y = f(x_1, x_2, \dots, x_k) + \varepsilon = \beta_0 + \sum_{i=1}^k \beta_i x_i + \sum_{i=1}^k \sum_{j=1}^k \beta_{ij} x_i x_j + \varepsilon \quad (6)$$

2.4. Previous work

Experimental design based upon principal properties for exploring different dimensions of the reaction space is discussed by Carlson [1]. Exploration of the reaction space using designs based on principal properties in combination with fractional factorial designs has been discussed by Carlson and Lundstedt [3].

We previously presented another method [4] for solving the problem, based on Fedorov's iterative algorithm [5]. The algorithm was based on the *D-optimality* criterion [1]. A similar approach to the design problem has been given by Martin *et al.* [6] in the context of designing combinatorial libraries.

The problem with the *D-optimality* approach is that the model matrix $\mathbf{X}$ has to have full column rank. In some cases this will not be the case, as we will show later, and therefore the *D-optimality* approach will fail. If methods that require full column rank are to be used, the model in Equation (6) might have to be modified. In a totally new experimental situation this is not desirable.

Another possible solution to the problem of *D-optimality* with a rank-deficient candidate model matrix $\mathbf{X}$, as has been suggested by one referee, would be the following procedure. First decompose $\mathbf{X}$ into principal components. Then use the corresponding PC scores as new variables and determine a *D-optimal* design by some exchange algorithm. Since algorithms for the construction of *D-optimal* designs necessitate the computation of variance functions of all possible candidate points as well as the computation of intermediary determinants, such a procedure will be computationally intensive, especially when the candidate matrix is large. The procedure outlined below avoids this problem and yields optimal designs that span the possible experimental variations without using extensive computations.

3. METHOD

In this section we will first give a mathematical interpretation of the problem described above and then present an algorithm that solves the problem.

3.1. Outline of strategy

We suggest that the design is constructed in two steps. The first step is a selection of reasonable test candidates from each class of constituents of the reaction space. This selection is made so that the range of variation of interest to the chemical problem is spanned by the selected items. The second step is to construct a matrix $\mathbf{D}$ of candidate experiments. This matrix will thus represent a multilevel full factorial design in the selected discrete settings of the variables of the reaction space. From $\mathbf{D}$ a candidate model matrix $\mathbf{C}$ is constructed by appending a column of ones, columns of the cross-products of the variables, and columns with squared variables included in the model. In the general case, $\mathbf{C}$ is very large.

In the following subsections we present how to select a small subset of M experiments from $\mathbf{C}$ so that the systematic variation in the reaction space can be explored, i.e. an experimental design. This selection will be denoted by the matrix $\mathbf{X}$.

3.2. Mathematical interpretation of the problem

Assuming we can describe the outcome of each experiment by a second-order Taylor expansion model as given in Equation (6), we can express the result of performing M experiments in matrix notation as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\epsilon} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k}^2 \\ 1 & x_{21} & x_{22} & \cdots & x_{2k}^2 \\ \vdots & \vdots & \ddots & & \vdots \\ 1 & x_{M1} & x_{M2} & \cdots & x_{Mk}^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{kk} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_M \end{bmatrix} \quad (7)$$

where $\mathbf{X}$ is as defined in the previous subsection. Now the best estimate of the coefficients $\mathbf{b}$ in the least-squares sense is

$$\hat{\mathbf{b}} = \mathbf{X}^\dagger \mathbf{y} \quad (8)$$

where $\mathbf{X}^\dagger$ is the pseudoinverse of $\mathbf{X}$, as given by *singular value decomposition* (SVD) [7]. If the matrix $\mathbf{X}$ has full column rank, the pseudoinverse simplifies to

$$\mathbf{X}^\dagger = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \quad (9)$$

which affords the ordinary least squares (OLS) estimate of $\hat{\mathbf{b}}$. It is of course also possible to fit the model by PLS. We have chosen the pseudoinverse to demonstrate the design procedure. An optimal design for fitting the model through the pseudoinverse method will also be a good design for fitting the model by PLS.

We now define the optimal matrix $\mathbf{X}$ as the matrix minimizing

$$J = E\left\{\|\hat{\mathbf{b}} - \mathbf{b}\|^2\right\} \quad (10)$$

where $\mathbf{X}$ is a matrix with M rows selected among the rows of the candidate matrix $\mathbf{C}$, so that the mean squared error of the model parameters $\hat{\mathbf{b}}$ is minimized:

$$E\left\{\|\hat{\mathbf{b}} - \mathbf{b}\|^2\right\} = E\left\{(\hat{\mathbf{b}} - \mathbf{b})^T (\hat{\mathbf{b}} - \mathbf{b})\right\} \quad (11)$$

Inserting (8) into (11) yields

$$E\left\{(\mathbf{X}^\dagger \mathbf{y} - \mathbf{b})^T (\mathbf{X}^\dagger \mathbf{y} - \mathbf{b})\right\} \quad (12)$$

Under the assumption that the experimental error $\boldsymbol{\epsilon}$ is independent of $\mathbf{X}$ and $E\{\boldsymbol{\epsilon}\} = 0$, we obtain

$$\begin{aligned} & E\left\{(\mathbf{X}^\dagger \mathbf{X}\mathbf{b} + \mathbf{X}^\dagger \boldsymbol{\epsilon} - \mathbf{b})^T (\mathbf{X}^\dagger \mathbf{X}\mathbf{b} + \mathbf{X}^\dagger \boldsymbol{\epsilon} - \mathbf{b})\right\} \\ &= E\left\{(\mathbf{b}^T \mathbf{X}^T \mathbf{X}^{\dagger T} + \boldsymbol{\epsilon}^T \mathbf{X}^{\dagger T} - \mathbf{b}^T)(\mathbf{X}^\dagger \mathbf{X}\mathbf{b} + \mathbf{X}^\dagger \boldsymbol{\epsilon} - \mathbf{b})\right\} \\ &= E\left\{\mathbf{b}^T \mathbf{X}^T \mathbf{X}^{\dagger T} \mathbf{X}^\dagger \mathbf{X}\mathbf{b} - \mathbf{b}^T \mathbf{X}^T \mathbf{X}^{\dagger T} \mathbf{b} + \boldsymbol{\epsilon}^T \mathbf{X}^{\dagger T} \mathbf{X}^\dagger \boldsymbol{\epsilon} - \mathbf{b}^T \mathbf{X}^\dagger \mathbf{X}\mathbf{b} + \mathbf{b}^T \mathbf{b}\right\} \end{aligned} \quad (13)$$

Since $\mathbf{X}^\dagger \mathbf{X}$ is a projection matrix $\mathbf{P}_\mathbf{X}$ onto the row space [8] of $\mathbf{X}$, Equation (13) simplifies to

$$\begin{aligned}
& E\{\mathbf{b}^T \mathbf{P}_X \mathbf{b} - \mathbf{b}^T \mathbf{P}_X \mathbf{b} + \boldsymbol{\epsilon}^T \mathbf{X}^{\dagger T} \mathbf{X}^{\dagger} \boldsymbol{\epsilon} - \mathbf{b}^T \mathbf{P}_X \mathbf{b} + \mathbf{b}^T \mathbf{b}\} \\
&= E\{\boldsymbol{\epsilon}^T \mathbf{X}^{\dagger T} \mathbf{X}^{\dagger} \boldsymbol{\epsilon} - \mathbf{b}^T \mathbf{P}_X \mathbf{b} + \mathbf{b}^T \mathbf{b}\} \\
&= E\{\boldsymbol{\epsilon}^T \mathbf{X}^{\dagger T} \mathbf{X}^{\dagger} \boldsymbol{\epsilon} + \mathbf{b}^T (\mathbf{I} - \mathbf{P}_X) \mathbf{b}\} \\
&= E\{\boldsymbol{\epsilon}^T \mathbf{X}^{\dagger T} \mathbf{X}^{\dagger} \boldsymbol{\epsilon} + \mathbf{b}^T \mathbf{P}_X^{\perp} \mathbf{b}\} \\
&= E\{\boldsymbol{\epsilon}^T \mathbf{X}^{\dagger T} \mathbf{X}^{\dagger} \boldsymbol{\epsilon}\} + \|\mathbf{P}_X^{\perp} \mathbf{b}\|^2
\end{aligned} \tag{14}$$

where $\mathbf{P}_X^{\perp}$ is the projection matrix that projects on the orthogonal complement of the row space of $\mathbf{X}$. This term decreases when the rank of $\mathbf{X}$ increases. Thus, if $\mathbf{X}$ has full column rank, the second term in (14) vanishes and the error reduces to

$$J = E\{\boldsymbol{\epsilon}^T \mathbf{X}^{\dagger T} \mathbf{X}^{\dagger} \boldsymbol{\epsilon}\} \tag{15}$$

This leads to the first step of our algorithm, where we select rows from a matrix of candidate experiments so that the rows of the experimental model matrix $\mathbf{X}$ are as orthogonal as possible. Doing this, $\mathbf{X}$ will have maximum possible rank and the second term in (14) is minimized. The second step in the algorithm then appends $\mathbf{X}$ with rows that minimize (15). This will not alter the value of the second term in Equation (14).

It can be shown that minimizing (15) is equivalent to minimizing

$$J = \sum_{k=1}^r \frac{1}{\sigma_k^2} \tag{16}$$

where σ_k^2 , $k = 1, 2, \dots, r$, are the eigenvalues of $\mathbf{X}^T \mathbf{X}$ and r is the rank of $\mathbf{X}$. The proof is given in Appendix II.

3.3. The algorithm

In this subsection we will show how to use SVD to select the experiments we want to perform. The algorithm consists of two steps. First we choose a number of rows equal to the rank of the candidate matrix. The second step then deals with selecting additional rows.

3.3.1. Step 1: creating a maximum rank matrix $\mathbf{X}$. Any $m \times n$ matrix $\mathbf{C}$ can be factored into

$$\mathbf{C} = \mathbf{U} \boldsymbol{\Sigma} \mathbf{V}^T \tag{17}$$

where the columns of $\mathbf{U}$ are eigenvectors of $\mathbf{C} \mathbf{C}^T$ and the columns of $\mathbf{V}$ are eigenvectors of $\mathbf{C}^T \mathbf{C}$. If the rank of $\mathbf{C}$ is r , the r singular values on the diagonal of $\boldsymbol{\Sigma}$ are the square roots of the non-zero eigenvalues of $\mathbf{C} \mathbf{C}^T$ and $\mathbf{C}^T \mathbf{C}$. Furthermore, if $\mathbf{C}$ has rank r , the first r columns of $\mathbf{V}$ form an orthogonal basis for the row space of $\mathbf{C}$, and the first r columns of $\mathbf{U}$ form an orthogonal basis for the column space of $\mathbf{C}$. Another useful property of the columns in $\mathbf{V}$ is that the first one points out the direction in which we have the largest variation in the row space, the next column points out the second most important direction, etc. If, using as few experiments as possible, we wish to span the row space of $\mathbf{C}$ and at the same time cover as much of the total variation as possible, selecting the experiments as the first r columns of $\mathbf{V}$ is optimal. In a real-life situation, however, it is not

likely that we can set up experiments that satisfy this. What we have to do instead is to select the row in $\mathbf{C}$ that is the most parallel to the first column in $\mathbf{V}$; that is, choose the row $\mathbf{c}_i^T$ in $\mathbf{C}$ that maximizes the absolute value of the scalar product $\mathbf{c}_i^T \mathbf{v}_1$.

When the first experiment has been chosen, we can remove the component in this direction from all rows in $\mathbf{C}$. The resulting matrix will have rank exactly one less than $\mathbf{C}$. Removing the projection onto row $\mathbf{c}_i^T$ from all other rows $\mathbf{c}_k^T$ is done according to

$$\hat{\mathbf{c}}_k = \mathbf{c}_k - \frac{\mathbf{c}_i^T \mathbf{c}_k}{\mathbf{c}_i^T \mathbf{c}_i} \mathbf{c}_i \quad (18)$$

Below we explain how this procedure then can be repeated to select as many rows as the rank r of the candidate matrix. The algorithm can be summarized in the following steps.

1. Let $\mathbf{C}_0$ be the original candidate matrix, consisting of all possible experiments.
2. Factor $\mathbf{C}_j$ (where $\mathbf{C}_j$ is the candidate matrix after j iterations) in $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$ using *singular value decomposition* (SVD).
3. Select the row $\mathbf{c}_i^T$ from $\mathbf{C}_j$ that is most parallel to the first column in $\mathbf{V}$, by finding the row that maximizes $\text{abs}(\mathbf{c}_i^T \mathbf{v}_1)$. Add the corresponding row of $\mathbf{C}_0$ to the model matrix $\mathbf{X}$.
4. Construct a new matrix $\mathbf{C}_{j+1}$ where the rows $\hat{\mathbf{c}}_k^T$ are defined by Equation (18).
5. Repeat from step 2 until the desired number of experiments has been selected (less than or equal to the rank of $\mathbf{C}_0$).

3.3.2. Step 2: adding rows to the design. The algorithm above yields a design which spans the reaction space. It is possible, however, to decrease the uncertainty in the model parameters by performing additional experiments. In this subsection we show how to do this so that (16) will be minimized.

To include additional rows, we select the row that increases the smallest singular value of $\mathbf{X}$ the most. This is achieved by selecting the row that has the largest component in the direction of $\mathbf{v}_r^T$, i.e. the singular vector corresponding to the smallest singular value. The reason for this is that no singular value can decrease if a matrix is appended with an additional row [9], and the sum in Equation (16) will be smallest if all σ_k^2 are equal. The implementation of this is very similar to the implementation of the first step of the algorithm. A more straightforward but much more computationally demanding way of appending the design is to test all possible candidate experiments and choose the one that decreases the sum in (16) the most. We have tested this, but did not find any significant improvement at all.

3.4. Design matrix

The actual experimental design is obtained from $\mathbf{X}$ by identifying the test items from their corresponding principal property scores.

4. CHEMICAL EXAMPLE

The acid-catalyzed rearrangement of phenyl hydrazones from ketones is known as the Fischer indole synthesis. Phenyl hydrazones derived from dissymmetric ketones can give rise to two regioisomeric indoles (see Figure 2). An extensive study to clarify the roles played by the properties of the catalysts and solvents for determining the regioselectivity of the reaction has been published by Prochazka and Carlson [10]. The problem to solve was to determine whether or not there exists a combination of

Figure 2. Fischer indole synthesis from dissymmetric ketones.

catalysts and solvents which can give a regioselective reaction. To illustrate the strategy suggested in this paper, we will use the data from the book by Carlson [1].

4.1. Data

The reaction space consists of ketones, Lewis acid catalysts and solvents. From the principal property projections shown in Figures 3–5, the test items summarized in Tables I–III were selected. For ketones, the ν parameter, described by Charton [11], was included to take possible effects of steric congestions into account.

We note that the sole purpose of this example is to illustrate our suggested strategy. We have chosen this example for four reasons

1. The test items used in the experiments were chosen by their principal properties.
2. The experiments were carried out by a single experimenter and the data are therefore consistent.
3. The data set contains a sufficiently large number of experiments.
4. The data were available.

The total number of combinations is 600 possible experiments. Of these, 296 were tested experimentally, and out of these, 162 experiments afforded the Fischer indole reaction. The response y is the regioisomeric excess, RE. These experiments are summarized in Appendix I.

Since we have used only the 'successful' experiments in the data set, we agree, at least in part, with the comment by one referee that this is not a true screening experiment. If, as in this case, we consider the specific question posed to the experimental system, i.e. 'Is there a combination of catalyst and

Figure 3. Principal property projections of ketones.

Table I. Principal property values for ketones used in experiment

Ketone (ID no.)	Principal properties		Steric parameter ν	
	t_1	t_2	R^1CH_2	R^2CH_2
3-Hexanone (50)	2.34	-0.16	0.68	0.56
2-Hexanone (49)	2.46	-0.15	0.52	0.68
3-Undecanone (78)	-0.52	2.16	0.68	0.56
1-Phenyl-2-butanone (72)	-0.74	0.42	0.56	0.70
5-Methyl-3-heptanone (59)	1.17	0.74	0.56	1.0

Figure 4. Principal property projections of Lewis acid catalysts.

Table II. Principal property values for Lewis acids used in experiment

Lewis acid (ID no.)	Principal properties	
	t_1	t_2
BF_3 (44)	7.05	2.80
$CuCl$ (13)	-3.42	1.33
ZnI_2 (99)	-2.25	-1.77
$TiCl_4$ (3)	4.35	-1.48
$ZnCl_2$ (15)	-1.11	0.79
$SbCl_5$ (27)	3.07	-3.18
PCl_3 (113)	1.32	-2.70
CuI (97)	-4.00	-0.47
$FeCl_3$ (10)	-0.06	-0.48
$SiCl_4$ (20)	3.18	-1.78
$AlCl_3$ (17)	0.97	1.83
$SnCl_4$ (23)	2.20	-2.22

Figure 5. Principal property projections of solvents.

solvent that will yield a regioselective Fischer indole reaction?', the data set is fully appropriate. There are of course a number of combinations of substrate, catalysts and solvents that do not give the Fischer indole reaction, but these combinations are of no interest for finding the answer to the question posed. If the question posed had been 'What combination of Lewis acid catalysts and solvents can be used in the Fischer indole reaction?', this would have been a totally different problem and of course the selected data set would have been inappropriate.

4.2. Computational results

In this subsection we will show that the present algorithm can be used to obtain good predictive models from fewer experiments than were used in the study by Prochazka and Carlson [11].

A candidate model matrix **C** was constructed from the experiments in Appendix I using the variables shown in Tables I–III and expanding these variables into columns with cross-products and squared variables. The effective rank of **C** was found to be 35, out of 44 columns. We have fitted a second-order polynomial model (Equation (6)) using a least squares estimator. The model has then been used to

Table III. Principal property values for solvents used in experiment

Solvent (ID no.)	Principal properties	
	t_1	t_2
Sulfolane (28)	3.29	2.21
Carbon disulfide (78)	−3.27	0.98
<i>N, N</i> -Dimethylacetamide (32)	1.81	0.15
Quinoline (51)	−0.29	3.13
1,2-Dichlorobenzene (57)	−0.99	2.27
Dimethyl sulfoxide (26)	2.54	0.95
Carbon tetrachloride (79)	−2.49	0.38
Chloroform (53)	−1.56	−0.59
Tetrahydrofuran (63)	−1.4	−1.57
Hexane (82)	−3.00	−1.20

Figure 6. Performance of proposed algorithm for example with 162 candidate experiments. The candidate model matrix was 162×44 and had rank 35. The variance of the prediction errors was used as a quality measure.

predict the outcome of all 162 experiments. The variance of the resulting prediction errors is plotted in Figure 6 as a function of the number of selected experiments used for the least squares fit.

We see from the figure that the prediction error variance drops rapidly once maximum column rank is reached. Performing more than around 50 experiments would be a waste of time.

We note that the algorithm produces a set of experiments ordered so that the first experiment represents the direction through the reaction space with the largest variance. The following experiments are ordered in decreasing order of how much of the remaining variance they describe. This is an advantage of the method, since the ordered set of experiments can be used in explorative experimental studies conducting one experiment at a time, when the objective is to find a useful combination of reagent, solvent, catalyst, etc. Since the experiments are chosen to be as orthogonal to each other as possible, each new experiment provides new and complementary information as was obtainable from the previous experiments. This allows for a systematic search of the efficient dimensions of the reaction space. It may well be that a satisfactory result is obtained early and in that case we do not have to run experiments to reach a maximum rank. If the objective is to examine in detail how the properties of the constituents influence the outcome, then of course we need good estimates of the model parameters. To achieve this, we see in Figure 6 that at least 35 experiments are needed in the present case.

5. DISCUSSION AND CONCLUSIONS

In this paper we have presented an algorithm for the construction of optimal experimental designs, for exploring the discrete variations defined by the reaction space. We note that this problem is identical to the problems encountered in combinatorial chemistry in drug research when the objective is to

construct target molecules from sets of discrete building blocks. Provided that these sets have been characterized by property descriptors, the presented algorithm can be used to design combinatorial libraries. These libraries will then efficiently cover the variation spanned by the building blocks.

In previous work the *D*-optimality criterion has been used for constructing optimal designs. When the candidate matrix does not have full column rank, the *D*-optimality approach will fail. This is the case in the presented example of the Fischer indole synthesis. It is possible, however, to obtain a good predictive model even when the model matrix does not have full column rank. The problem of model fitting with rank-deficient **X** matrices can be solved by using PLS, as was pointed out by one referee. We have used the pseudoinverse to fit the models. However, the method used to fit the model does not solve the design problem.

In the previous section we show that our algorithm can handle this case with very good result. In an experimental situation where the chemist removes certain combinations of the constituents, since he immediately realizes that these combinations will not work, the effective rank of the candidate matrix might be affected. This is similar to the case in the example presented in this paper, where we have used only those experiments which actually afforded the Fischer indole reaction. It was suggested by one referee that it would be appropriate to scale and center the variables in the candidate matrix and to center the variables of the candidate model matrix. We agree that this facilitates the interpretation of the estimated model. However, it will not change the efficient rank of the matrix and hence not the efficient dimensions of the reaction space necessary to explore. There is, however, another aspect related to scaling that deserves a comment. The basis of the proposed design strategy is the characterization of the constituents of the reaction space by their corresponding principal properties. When these are determined, the original matrix of chemically relevant descriptors is scaled and centered prior to the principal component analysis. The scaling of the chemical information is actually made at this point. When the particular test items are then selected from the corresponding principal property score projections, the chemist makes a selection that spans the range of variation of interest. This means that selected items showing a large range of variation do so because the chemist is interested in a large variation in these principal properties. It would therefore be inappropriate to scale away this important chemical background information.

The presented algorithm is fast, numerically stable and easily implemented on a standard PC. The example presented in this paper was processed with a MATLAB implementation of the algorithm on a 200 MHz Pentium-based computer. The execution time was less than 30 s.

Below is a summary of the proposed strategy

1. Analyze the problem and determine the classes of constituents defining the reaction space. Determine which response(s) should be measured.
2. Define pertinent descriptors for these classes. Compile descriptor data for possible test items of each class. Determine the principal properties.
3. Use the score plots of the principal properties to select subsets of test items from each class so that the desired spread of the molecular properties is covered by the selected test items.
4. Determine the candidate design matrix as a full factorial design in the discrete settings defined by the selected test items. Transform this matrix into a matrix where candidates are replaced by the corresponding principal property scores.
5. Assign a tentative Taylor expansion model for the screening.
6. Compute the candidate model matrix **C** by appending columns corresponding to terms in the model (constant, cross-product and square terms).
7. Use the presented algorithm to select experiments that span the row space of **C**.
8. If desired, use the second step of the presented algorithm to select additional experiments to improve the model fit.

9. Run the experiments and fit the model. Assess the significance of the variables.
10. Confirm conclusions by experiments!

The above strategy yields an experimental design by selecting one experiment at a time. Step 7 selects experiments that span the row space of C , i.e. the significant dimensions of the reaction space. Since we select experiment vectors that are as parallel as possible to the singular vectors, this procedure peels off one dimension at a time of the reaction space in a sequence of decreasing variance. If there are experiments that are perfectly parallel to the singular vectors, this procedure will yield a perfectly D -optimal design that spans the row space. However, it is not likely that there are experiments that are perfectly parallel to the singular vectors. The present algorithm yields a design that is the best approximate choice. In step 7 the number of experiments is augmented by selecting complementary experiments in those directions through the reaction space in which the information given by the previous experiments is feeble. This improves the condition number of the design and optimizes the design toward the criterion given in Equation (16). Owing to the sequential nature of the design construction procedure, we cannot claim that this strategy yields the ultimately best design available for estimating the model parameters (e.g. according to the D -optimality criterion). Nevertheless, we claim that a design constructed according to the above principles is good enough for the purpose discussed in the present paper.

ACKNOWLEDGEMENTS

Financial support from the *Norwegian Research Council* is gratefully acknowledged. The authors would also like to thank Mr Paul Petersen for linguistic assistance.

APPENDIX I. Candidate reaction systems in the Fischer indole synthesis

No.	Reaction system			RE
	Ketone	Lewis acid	Solvent	
1	3-Hexanone	ZnI ₂	Sulfolane	57.4
2	3-Hexanone	TiCl ₄	Sulfolane	48.8
3	3-Hexanone	ZnCl ₂	Sulfolane	52.4
4	3-Hexanone	PCl ₃	Sulfolane	51.2
5	3-Hexanone	FeCl ₃	Sulfolane	58.8
6	3-Hexanone	AlCl ₃	Sulfolane	63.2
7	3-Hexanone	BF ₃	Carbon disulfide	50.6
8	3-Hexanone	ZnI ₂	Carbon disulfide	43.6
9	3-Hexanone	TiCl ₄	Carbon disulfide	54.6
10	3-Hexanone	ZnCl ₂	Carbon disulfide	44.0
11	3-Hexanone	SbCl ₅	Carbon disulfide	54.0
12	3-Hexanone	FeCl ₃	Carbon disulfide	61.0
13	3-Hexanone	AlCl ₃	Carbon disulfide	60.4
14	3-Hexanone	BF ₃	1,2-Dichlorobenzene	36.2
15	3-Hexanone	ZnI ₂	1,2-Dichlorobenzene	42.0
16	3-Hexanone	TiCl ₄	1,2-Dichlorobenzene	22.0
17	3-Hexanone	ZnCl ₂	1,2-Dichlorobenzene	39.0
18	3-Hexanone	PCl ₃	1,2-Dichlorobenzene	52.4
19	3-Hexanone	FeCl ₃	1,2-Dichlorobenzene	54.8
20	3-Hexanone	AlCl ₃	1,2-Dichlorobenzene	42.0
21	3-Hexanone	FeCl ₃	Dimethyl sulfoxide	58.0
22	3-Hexanone	BF ₃	Tetrahydrofuran	57.4
23	3-Hexanone	TiCl ₄	Tetrahydrofuran	58.6

APPENDIX I. Candidate reaction systems in the Fischer indole synthesis (*continued*)

No.	Reaction system			RE
	Ketone	Lewis acid	Solvent	
24	3-Hexanone	FeCl ₃	Tetrahydrofuran	58·0
25	3-Hexanone	AlCl ₃	Tetrahydrofuran	58·0
26	2-Hexanone	BF ₃	Sulfolane	100·0
27	2-Hexanone	ZnI ₂	Sulfolane	100·0
28	2-Hexanone	TiCl ₄	Sulfolane	100·0
29	2-Hexanone	ZnCl ₂	Sulfolane	100·0
30	2-Hexanone	PCl ₃	Sulfolane	100·0
31	2-Hexanone	FeCl ₃	Sulfolane	100·0
32	2-Hexanone	AlCl ₃	Sulfolane	100·0
33	2-Hexanone	BF ₃	Carbon disulfide	100·0
34	2-Hexanone	ZnI ₂	Carbon disulfide	100·0
35	2-Hexanone	ZnCl ₂	Carbon disulfide	100·0
36	2-Hexanone	BF ₃	1,2-Dichlorobenzene	98·0
37	2-Hexanone	ZnI ₂	1,2-Dichlorobenzene	100·0
38	2-Hexanone	TiCl ₄	1,2-Dichlorobenzene	100·0
39	2-Hexanone	ZnCl ₂	1,2-Dichlorobenzene	98·0
40	2-Hexanone	PCl ₃	1,2-Dichlorobenzene	100·0
41	2-Hexanone	FeCl ₃	1,2-Dichlorobenzene	100·0
42	2-Hexanone	AlCl ₃	1,2-Dichlorobenzene	100·0
43	2-Hexanone	BF ₃	Tetrahydrofuran	90·0
44	2-Hexanone	TiCl ₄	Tetrahydrofuran	84·0
45	2-Hexanone	PCl ₃	Tetrahydrofuran	88·0
46	2-Hexanone	FeCl ₃	Tetrahydrofuran	92·0
47	2-Hexanone	AlCl ₃	Tetrahydrofuran	98·0
48	3-Undecanone	BF ₃	Sulfolane	23·6
49	3-Undecanone	ZnI ₂	Sulfolane	32·0
50	3-Undecanone	TiCl ₄	Sulfolane	30·0
51	3-Undecanone	ZnCl ₂	Sulfolane	30·0
52	3-Undecanone	PCl ₃	Sulfolane	31·0
53	3-Undecanone	FeCl ₃	Sulfolane	36·0
54	3-Undecanone	AlCl ₃	Sulfolane	36·0
55	3-Undecanone	BF ₃	Carbon disulfide	21·0
56	3-Undecanone	ZnI ₂	Carbon disulfide	34·2
57	3-Undecanone	PCl ₃	Carbon disulfide	33·0
58	3-Undecanone	AlCl ₃	Carbon disulfide	30·0
59	3-Undecanone	BF ₃	1,2-Dichlorobenzene	23·0
60	3-Undecanone	ZnI ₂	1,2-Dichlorobenzene	30·4
61	3-Undecanone	ZnCl ₂	1,2-Dichlorobenzene	29·0
62	3-Undecanone	PCl ₃	1,2-Dichlorobenzene	31·2
63	3-Undecanone	FeCl ₃	1,2-Dichlorobenzene	34·0
64	3-Undecanone	AlCl ₃	1,2-Dichlorobenzene	28·0
65	3-Undecanone	BF ₃	Dimethyl sulfoxide	32·2
66	3-Undecanone	ZnI ₂	Dimethyl sulfoxide	2·6
67	3-Undecanone	ZnCl ₂	Dimethyl sulfoxide	14·0
68	3-Undecanone	FeCl ₃	Dimethyl sulfoxide	22·0
69	3-Undecanone	BF ₃	Tetrahydrofuran	35·0
70	3-Undecanone	ZnI ₂	Tetrahydrofuran	12·0
71	3-Undecanone	TiCl ₄	Tetrahydrofuran	40·0
72	3-Undecanone	PCl ₃	Tetrahydrofuran	33·0
73	3-Undecanone	FeCl ₃	Tetrahydrofuran	34·0
74	3-Undecanone	AlCl ₃	Tetrahydrofuran	38·0
75	1-Phenyl-2-butanone	BF ₃	Sulfolane	68·0

APPENDIX I. Candidate reaction systems in the Fischer indole synthesis (*continued*)

No.	Reaction system			RE
	Ketone	Lewis acid	Solvent	
76	1-Phenyl-2-butanone	ZnI ₂	Sulfolane	44.0
77	1-Phenyl-2-butanone	TiCl ₄	Sulfolane	56.0
78	1-Phenyl-2-butanone	ZnCl ₂	Sulfolane	38.0
79	1-Phenyl-2-butanone	PCl ₃	Sulfolane	26.0
80	1-Phenyl-2-butanone	FeCl ₃	Sulfolane	46.0
81	1-Phenyl-2-butanone	AlCl ₃	Sulfolane	28.0
82	1-Phenyl-2-butanone	BF ₃	Carbon disulfide	88.0
83	1-Phenyl-2-butanone	ZnI ₂	Carbon disulfide	58.0
85	1-Phenyl-2-butanone	ZnCl ₂	Carbon disulfide	70.0
86	1-Phenyl-2-butanone	PCl ₃	Carbon disulfide	66.0
87	1-Phenyl-2-butanone	FeCl ₃	Carbon disulfide	28.0
88	1-Phenyl-2-butanone	AlCl ₃	Carbon disulfide	88.0
89	1-Phenyl-2-butanone	BF ₃	1,2-Dichlorobenzene	62.0
90	1-Phenyl-2-butanone	ZnI ₂	1,2-Dichlorobenzene	54.0
91	1-Phenyl-2-butanone	TiCl ₄	1,2-Dichlorobenzene	62.0
92	1-Phenyl-2-butanone	ZnCl ₂	1,2-Dichlorobenzene	46.0
93	1-Phenyl-2-butanone	PCl ₃	1,2-Dichlorobenzene	52.0
94	1-Phenyl-2-butanone	FeCl ₃	1,2-Dichlorobenzene	90.0
95	1-Phenyl-2-butanone	AlCl ₃	1,2-Dichlorobenzene	42.0
96	1-Phenyl-2-butanone	BF ₃	Dimethyl sulfoxide	50.0
97	1-Phenyl-2-butanone	FeCl ₃	Dimethyl sulfoxide	66.0
98	1-Phenyl-2-butanone	BF ₃	Tetrahydrofuran	60.0
99	1-Phenyl-2-butanone	ZnI ₂	Tetrahydrofuran	88.0
100	1-Phenyl-2-butanone	TiCl ₄	Tetrahydrofuran	76.0
101	1-Phenyl-2-butanone	PCl ₃	Tetrahydrofuran	38.0
102	1-Phenyl-2-butanone	FeCl ₃	Tetrahydrofuran	56.0
103	1-Phenyl-2-butanone	AlCl ₃	Tetrahydrofuran	50.0
104	5-Methyl-3-heptanone	BF ₃	Sulfolane	100.0
105	5-Methyl-3-heptanone	ZnI ₂	Sulfolane	100.0
106	5-Methyl-3-heptanone	TiCl ₄	Sulfolane	100.0
107	5-Methyl-3-heptanone	ZnCl ₂	Sulfolane	100.0
108	5-Methyl-3-heptanone	PCl ₃	Sulfolane	100.0
109	5-Methyl-3-heptanone	FeCl ₃	Sulfolane	100.0
110	5-Methyl-3-heptanone	AlCl ₃	Sulfolane	100.0
111	5-Methyl-3-heptanone	BF ₃	Carbon disulfide	100.0
112	5-Methyl-3-heptanone	ZnI ₂	Carbon disulfide	100.0
113	5-Methyl-3-heptanone	BF ₃	1,2-Dichlorobenzene	100.0
114	5-Methyl-3-heptanone	ZnI ₂	1,2-Dichlorobenzene	100.0
115	5-Methyl-3-heptanone	TiCl ₄	1,2-Dichlorobenzene	100.0
116	5-Methyl-3-heptanone	ZnCl ₂	1,2-Dichlorobenzene	100.0
117	5-Methyl-3-heptanone	PCl ₃	1,2-Dichlorobenzene	100.0
118	5-Methyl-3-heptanone	FeCl ₃	1,2-Dichlorobenzene	100.0
119	5-Methyl-3-heptanone	BF ₃	Dimethyl sulfoxide	100.0
120	5-Methyl-3-heptanone	FeCl ₃	Dimethyl sulfoxide	100.0
121	5-Methyl-3-heptanone	BF ₃	Tetrahydrofuran	100.0
122	5-Methyl-3-heptanone	TiCl ₄	Tetrahydrofuran	100.0
123	5-Methyl-3-heptanone	PCl ₃	Tetrahydrofuran	100.0
124	5-Methyl-3-heptanone	AlCl ₃	Tetrahydrofuran	100.0
125	3-Hexanone	TiCl ₄	Quinoline	34.0
126	3-Hexanone	SiCl ₄	Quinoline	56.4
127	3-Hexanone	BF ₃	Carbon tetrachloride	31.8
128	3-Hexanone	CuCl	Carbon tetrachloride	43.2

APPENDIX I. Candidate reaction systems in the Fischer indole synthesis (*continued*)

No.	Reaction system			RE
	Ketone	Lewis acid	Solvent	
129	3-Hexanone	ZnI ₂	Carbon tetrachloride	48.2
130	3-Hexanone	ZnCl ₂	Carbon tetrachloride	44.2
131	3-Hexanone	PCl ₃	Carbon tetrachloride	60.0
132	3-Hexanone	SiCl ₄	Carbon tetrachloride	56.8
133	3-Hexanone	BF ₃	Chloroform	43.2
134	3-Hexanone	ZnI ₂	Chloroform	36.0
135	3-Hexanone	TiCl ₄	Chloroform	36.0
136	3-Hexanone	ZnCl ₂	Chloroform	44.2
137	3-Hexanone	PCl ₃	Chloroform	57.2
138	3-Hexanone	FeCl ₃	Chloroform	53.4
139	3-Hexanone	SiCl ₄	Chloroform	42.8
140	3-Hexanone	AlCl ₃	Chloroform	45.8
141	3-Hexanone	SnCl ₄	Chloroform	33.0
142	3-Hexanone	SbCl ₅	Tetrahydrofuran	66.0
143	3-Hexanone	SnCl ₄	Tetrahydrofuran	62.0
144	2-hexanone	SbCl ₅	Tetrahydrofuran	100.0
145	2-hexanone	SnCl ₄	Tetrahydrofuran	88.0
146	2-hexanone	SiCl ₄	Sulfolane	100.0
147	2-hexanone	SnCl ₄	Sulfolane	100.0
148	3-Undecanone	BF ₃	Hexane	22.0
149	3-Undecanone	ZnI ₂	Hexane	28.0
150	3-Undecanone	ZnCl ₂	Hexane	14.0
151	3-Undecanone	PCl ₃	Hexane	32.0
152	3-Undecanone	FeCl ₃	Hexane	16.0
153	1-Phenyl-2-butanone	BF ₃	Hexane	70.0
154	1-Phenyl-2-butanone	ZnI ₂	Hexane	66.0
155	1-Phenyl-2-butanone	TiCl ₄	Hexane	56.0
156	1-Phenyl-2-butanone	PCl ₃	Hexane	48.0
157	1-Phenyl-2-butanone	FeCl ₃	Hexane	66.0
158	5-Methyl-3-heptanone	SiCl ₄	Sulfolane	100.0
159	5-Methyl-3-heptanone	SnCl ₄	Sulfolane	100.0
160	5-Methyl-3-heptanone	SiCl ₄	1,2-Dichlorobenzene	100.0
161	5-Methyl-3-heptanone	SiCl ₄	Tetrahydrofuran	100.0
162	5-Methyl-3-heptanone	SnCl ₄	Tetrahydrofuran	100.0

APPENDIX II. MINIMIZING $E \{ \boldsymbol{\epsilon}^T \mathbf{X}^{\dagger T} \mathbf{X}^{\dagger} \boldsymbol{\epsilon} \}$

In this appendix we show that, minimizing

$$J = E \{ \boldsymbol{\epsilon}^T \mathbf{X}^{\dagger T} \mathbf{X}^{\dagger} \boldsymbol{\epsilon} \} \quad (19)$$

is equivalent to minimizing

$$\sum_{k=1}^r \frac{1}{\sigma_k^2} \quad (20)$$

where σ_k^2 are the eigenvalues of $\mathbf{X}^T \mathbf{X}$ and r is the rank of $\mathbf{X}$. Since the expression inside the brackets is

a scalar, Equation (19) can be written as

$$J = E\{\text{tr}(\boldsymbol{\epsilon}^T \mathbf{X}^{\dagger T} \mathbf{X}^{\dagger} \boldsymbol{\epsilon})\} \quad (21)$$

Noting that $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$, this can be rewritten as

$$J = \text{tr}(\mathbf{X}^{\dagger} E\{\boldsymbol{\epsilon} \boldsymbol{\epsilon}^T\} \mathbf{X}^{\dagger T}) \quad (22)$$

We define the noise covariance matrix $\boldsymbol{\theta}$ as

$$\boldsymbol{\theta} = E\{\boldsymbol{\epsilon} \boldsymbol{\epsilon}^T\} \quad (23)$$

which leads to

$$J = \text{tr}(\mathbf{X}^{\dagger} \boldsymbol{\theta} \mathbf{X}^{\dagger T}) \quad (24)$$

Now $\mathbf{X}$ is factored using singular value decomposition, leading to

$$\mathbf{X} = \mathbf{U} \boldsymbol{\Sigma} \mathbf{V}^T \quad (25)$$

The pseudoinverse $\mathbf{X}^{\dagger}$ is then given by

$$\mathbf{X}^{\dagger} = \mathbf{V} \boldsymbol{\Sigma}^{\dagger} \mathbf{U}^T \quad (26)$$

where the diagonal of $\boldsymbol{\Sigma}^{\dagger}$ ($n \times m$) contains the reciprocals $1/\sigma_1, 1/\sigma_2, \dots, 1/\sigma_r$ of the non-zero diagonal elements of $\boldsymbol{\Sigma}$. All other elements of $\boldsymbol{\Sigma}^{\dagger}$ are zero. Inserting this into Equation (24), we obtain

$$J = \text{tr}(\mathbf{V} \boldsymbol{\Sigma}^{\dagger} \mathbf{U}^T \boldsymbol{\theta} \mathbf{U} \boldsymbol{\Sigma}^{\dagger T} \mathbf{V}^T) \quad (27)$$

Considering the special case when $\boldsymbol{\theta} = \mathbf{I} \eta^2$, this reduces to

$$J = \text{tr}(\mathbf{V} \boldsymbol{\Sigma}^{\dagger} \boldsymbol{\Sigma}^{\dagger T} \mathbf{V}^T) \eta^2 \quad (28)$$

Since the trace operator is invariant to a change of basis, this is equivalent to

$$J = \text{tr}(\boldsymbol{\Sigma}^{\dagger} \boldsymbol{\Sigma}^{\dagger T}) \eta^2 = \eta^2 \sum_{k=1}^r \frac{1}{\sigma_k^2} \quad (29)$$

which leads to the criterion that the trace of $\boldsymbol{\Sigma}^{\dagger} \boldsymbol{\Sigma}^{\dagger T}$ should be minimized, i.e.

$$\min_{\mathbf{x}} \sum_{k=1}^r \frac{1}{\sigma_k^2} \quad (30)$$

where σ_k^2 are the non-zero eigenvalues of $\mathbf{X}^T \mathbf{X}$ and r is the rank of $\mathbf{X}$.

This concludes the proof.

REFERENCES

1. Carlson R. *Design and Optimization in Organic Synthesis*. Elsevier: Amsterdam, 1992.
2. Wold S. Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics* 1978; **20**: 397–405.
3. Carlson R, Lundstedt T. Scope of organic synthetic reactions. Multivariate methods for exploring the reaction space. An example with the Wilgerodt–Kindler reaction. *Acta Chem. Scand. B* 1987; **41**: 164–173.
4. Carlson R, Carlson J. Strategy for screening discrete variations in organic synthesis. In *Proceedings of Chimométrie 97*. Société de Chimie Industrielle: Lyon, 1997; C11-1–C11-12.
5. Fedorov V. *Theory of Optimal Experiments*. Academic Press: New York, 1972.
6. Martin EJ, Blaney JM, Siani MA, Spellmeier DC, Wong AK, Moos WH. Measuring diversity: experimental design of combinatorial libraries for drug discovery. *J. Med. Chem.* 1995; **38**: 1431–1436.
7. Strang G. *Linear Algebra and Its Applications*. Academic Press: New York, 1980.
8. Golub GH, van Loan CF. *Matrix Computations*. Johns Hopkins University Press: Baltimore, MD, 1995.
9. Björck Å. *Numerical Methods for Least Squares Problems*. SIAM: Philadelphia, PA, 1996.
10. Prochazka MP, Carlson R. On the role of Lewis acid catalysts and solvents in the Fischer indole synthesis. *Acta Chem. Scand.* 1989; **5**: 651–659.
11. Charton M. Steric effects in drug design. The upsilon steric parameter—definition and determination. *Top. Curr. Chem.* 1982; **117**: 57–91.