Homepage of Chemometrics

Editorial
August 2002

http://www.acc.umu.se/~tnkjtg/Chemometrics/Editorial

OUTLINE

List of previous HoC Editorials are given on last page.

Introduction to Statistical Experimental Design - What is it? Why and Where is it Useful?

Johan Trygg & Svante Wold
University of Queensland, Australia & Umeå University, Sweden

Definition: Statistical experimental design, a.k.a. design of experiments (DoE) is the methodology of how to conduct and plan experiments in order to extract the maximum amount of information in the fewest number of runs.

1. Introduction

…We have a large reservoir of engineers (and scientists) with a vast background of engineering know- how. They need to learn statistical methods that can tap into the knowledge. Statistics used as a catalyst to engineering creation will, I believe, always result in the fastest and most economical progress… George Box, 1992

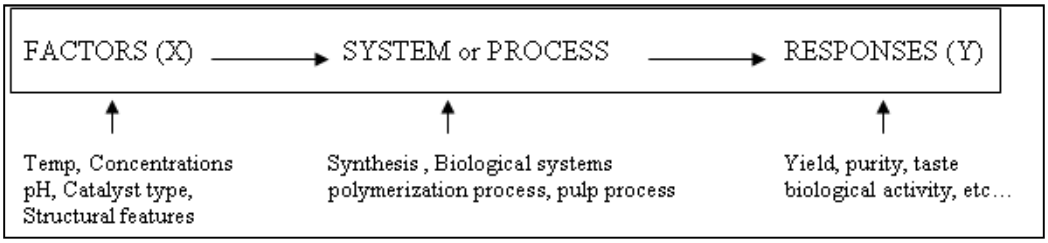
How shall I find the optimum? This is a common question everywhere in business. In research and development, often half of the resources are spent on solving optimization problems. With the rapidly rising costs of making experiments, it is essential that the optimization is done with as few experiments as possible. This is one important reason why statistical experimental design is needed.

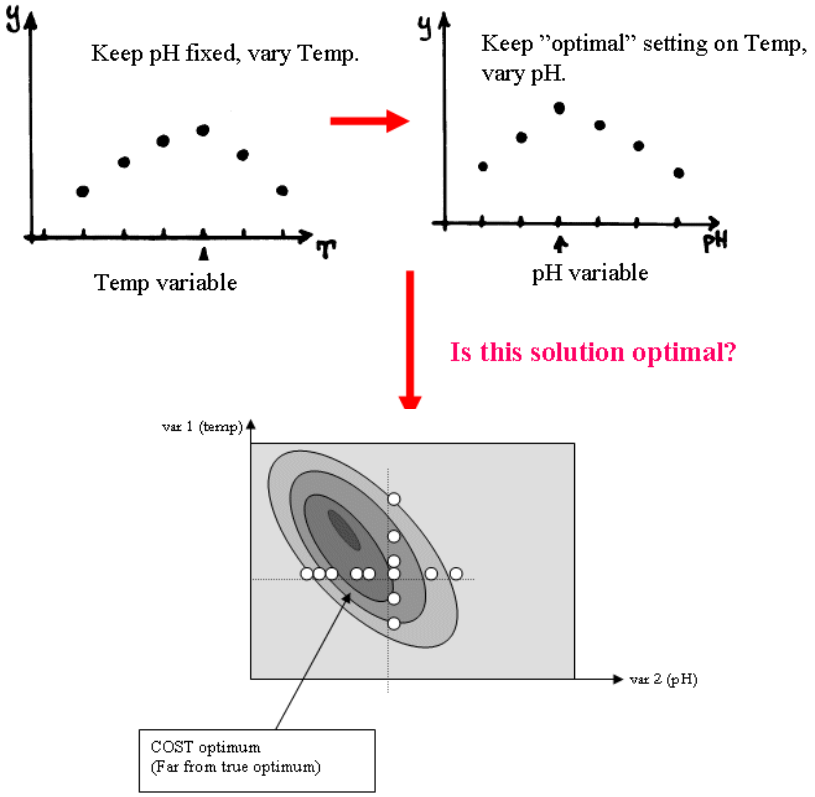
DoE originated in the 1920's by a British scientist, Sir R. A. Fisher, as a method to maximize the knowledge gained from experimental data and it has evolved over the last 70 years. Most experimentation involves several factors and are conducted in order to optimize processes and or investigate and understand the relationships between the factors and the characteristics of the process (reponses) of interest.

1.1 The traditional way - the COST approach [COST: Change One Separate factor at a Time]

Most experimentation today is done by changing levels of one factor (variable) at a time in an unsystematic way in order to try and find the optimum conditions of a complex system. Is this a good (efficient, rational, economic) strategy ? No!!
As shown by Fisher around 1925, changing one separate factor at a time (COST) does not give any information about the position of the optimum in the common case where there are interactions between factors. Then the COST approach gets stuck, usually far from the real optimum. However, the experimenter perceives that the optimum has been reached because changing one factor at a time does not lead to any further improvement. The COST approach is said to be pseudo-convergent.

What are the problems associated with the COST approach:
1. Does not lead to real optimum
2. Inefficient, unnecessarily many runs
3. Provides no information about what happens when factors are varied simultaneously (ignores interactions)
4. Provides less information about the variability





of the response
5. Isolated, unconnected experiments
6. Slow growth of knowledge, no mapping of experimental space.

Any measurement and experiment is influenced by noise. Under stable conditions, any process varies around its mean +/- 3 std dev. Two COST experiments may give two different results, but with no estimate (poor) of noise level. BUT by making a set of well planned experiments with correct analysis (DoE), we can separate "real effects" from noise and draw correct conclusions and act correctly. This means decreased variability and quality improvement. So to investigate systems involving several factors in presence of variability or noise one needs a better strategy than that based on changing one separate factor at a time and that strategy is given by experimental design (DoE).

1.2 What to do instead - Design of Experiments (DoE)

- Which factors have a real influence on the response?
- What are the best settings of the factors to achieve optimal conditions for best performance on a system?
- What are the predicted values of the responses for given settings of the factors in a model?

In 1925 Fisher started the development of methods of statistical experimental design. These methods have been further refined by Yule, Box, Stu and Bill Hunter, Scheffe, Cox, Taguchi, and others, so that today they comprise a tool box for virtually any optimization problem. The basic idea is to devise a small set of experiments, in which all pertinent factors are varied systematically. This set usually does not include more than ten to twenty experiments. The subsequent analysis of the resulting experimental data will identify the optimal conditions, the factors that most influence the results and those that do not, the presence of interactions and synergisms, and so on. The most important aspect of statistical experimental designs is that they provide a strict mathematical framework for changing all pertinent factors simultaneously, and achieve this in a small number of experimental runs. Most of us can only grasp the effect of one factor at a time in our minds, and that leads to the inefficient COST approach. We need the mathematics (and the computer) to keep track of the factors and their combinations.

- All factors are varied together over a set of experimental runs
- Noise is decreased by means of averaging
- The functional space is efficiently mapped, interactions and synergisms are seen

Typical examples when design of experiments (DoE) is useful involve the development of new products and processes, e.g. optimizing the quality and performance of an existing product or optimizing existing manufacturing processes of chemicals, polymers, materials, drugs and pharmaceuticals, foods and beverages, cosmetics, paints and so on.

2. Stages in the DoE process

2.1 Familiarization, fiddle around

Fiddling around a little. Problem formulation is extremely important during the whole process. Formulate question(s) stating the objectives and goals of the investigation. What do I want?

- What is the purpose?
- What are the objectives?
- Identify factors, factor ranges and types of factors (quantitative or qualitative)
- What is possible, experimentally, financially, environmentally?

If these objectives can not be put into words, there is no reason to continue because it shows that the investigators don't know what to do.

2.2 Screening (many factors)

Finding out a little about many factors. Which factors are the dominating ones. To assure that uncontrolled factors (humidity, etc..) do not bias the results, perform the runs in random order
Screening experiments give information about

- What are the important factors
- If we are in the correct region, (ranges)
- If there is curvature and if it masks the effects
- What to do next

Pareto's principle states that 20 % of the data (factors) account for 80 % of the information. Screening designs provide simple models with information about dominating variables, and information about ranges. In addition, they provide few experiments / factor which means that relevant information is gained in only a few experiments.
Linear models and interaction models are sufficient, since we are only interested in the effects. We merely ask, if a factor does influence the response, not how.

Linear model:	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \varepsilon$
Interaction model:	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \ldots + \varepsilon$

2.2.1 Screening designs:

Different types of screening designs exist, which one to choose depends on the problem. The most common one is the fractional factorials design.

Factorial designs form the basis for all classical experimental designs, both screening and RSM. For screening, we will concentrate on two-level designs. They are sufficient to estimate linear, and interaction models, and they require a very low number of experimental runs. There are some clear obvious advantages of using a factorial design.

- Averages are more stable than single observations.
- The more data one averages, the more reliable is the result.

The "equal opportunity" strategy (screening)

1. The factors are selected that a priori are believed to be the most influential on the response
2. A range is chosen for each factor
3. A (full factorial) or fractional factorial design (plus 3 ctr. points) is selected. This provide orthogonal, balanced estimates of the effects (and possibly interactions) with equal variance.
4. The experimental runs are performed in random order. This results in a ranking of the factors (and possibly interactions) with significance estimates.

Let us first describe the full factorial design, and later explain the fractional factorial design.

2.2.2 Full factorial design

The full factorial design is a set of experimental runs where every level of a factor is investigated at both levels of all the other factors. It is a balanced (orthogonal) design. A balanced design allows the estimation of a factor effect independently of all the other effects

1. $N=2^k$(+3 centerpoints) number of runs for *k* factors
2. Effects (main effects, interactions) are calculated using least squares (or Yates algorithm)
3. Estimate of noise is used for confidence interval for the effects, or hypothesis testing
4. When no estimate of noise is available, normal probability plot of the effects is used.
5. Noise can be estimated from
   -Replicated center points
   -Other replication
   -Residuals (higher level interactions)
   -Previous knowledge

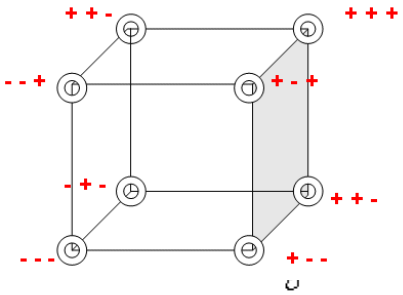


Figure. Example of a full factorial design in three variables, 2^3= 8 experiments

Notation
The low level of a factor will be denoted [-1] = [-]
The high level of a factor will be denoted [+1] =[+]
The center level of a factor will be denoted [0] = [0]

2.2.3 Fractional factorial design

Investigating more than 5 factors with the full factorial design becomes time consuming, 25=32, 26=64,27=128, etc. Instead, performing a fractional factorial design reduces that number quickly without the loss of too much information regarding the estimation of factors involved. Fractional factorial design takes advantage of the fact that 3-way and higher interactions are seldom significant. The downside, of course, for not performing all experiments is that confounding patterns are present. In other words, the estimated effects are not "pure" but instead mixed with higher degree interaction effects. This loss of information is the prize we need to pay

$$N = \frac{2^k}{2^p} + 3 = 2^{k-p} + 3 \text{ centerpoints}$$

for the reduction of the number of experiments.

- Fraction (k-p) selected by choice of generator
- Degree of confounding of coefficients can be derived from the defining relation The shortest word in the defining relation is the resolution of that fractional factorial design.

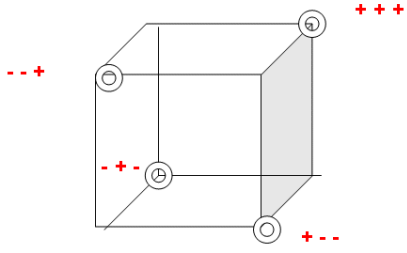


Figure. Example of a fractional factorial design in three variables, 2^{3-1}= 4 experiments

2.2.4 Placket-Burman

See http://www.itl.nist.gov/div898/handbook/pri/section3/pri335.htm for more info.

2.2.5 Special designs

In addition, special designs are needed in *constrained regions* or *mixture problems*. Mixture problems are common in the chemical, food and beverage, cosmetics, and drug industries. One reason is because the factors add up to 100 percent. This introduces a constraint on the design and must be handled with special tools and models. D-optimal designs or other special designs are used when there are constraints put on the factors for economical / synthetical / environmental or other reasons. The simpler factorial designs can not be used in such cases.

- D-optimal (and other "optimal" designs)
  - See http://www.itl.nist.gov/div898/handbook/pri/section5/pri521.htm for more info.
- Mixture designs
  - See http://www.itl.nist.gov/div898/handbook/pri/section5/pri54.htm for more info

2.3 Finding optimal region of operability

Simplex designs (see http://www.multisimplex.com/algorithm.htm for more info) and **steepest ascent** approaches (see http://www.itl.nist.gov/div898/handbook/pri/section5/pri5311.htm for more info.) are used to achieve optimal conditions in problems where experiments can only be done one at a time in a sequence or after the screening stage where one is often interested into moving the experimental region to its optimum.

2.3 Response surface modeling

and optimization (few factors)

After screening, the goal of the investigation is usually to create a valid map of the experimental domain (local space) given by the significant factors and their ranges. This is done with a quadratic polynomial model. The higher order models has an increased complexity and therefore also requires more experiments / factor than screening designs.
Different types of RSM designs

- Three level factorial designs
  - See http://www.itl.nist.gov/div898/handbook/pri/section3/pri339.htm for more info.
- Central composite designs (CCD)
  - See http://www.itl.nist.gov/div898/handbook/pri/section3/pri3361.htm for more info.
- Box Behnken designs
  - See http://www.itl.nist.gov/div898/handbook/pri/section3/pri3362.htm for more info.
- D-optimal designs
  - See http://www.itl.nist.gov/div898/handbook/pri/section5/pri521.htm for more info.

2.5 Robustness testing [read more in Ref 8]

In robustness testing of, for instance, an analytical method, the aim is to explore how sensitive the responses are to small changes in the factor settings. Ideally, a robustness test should show that the responses are not sensitive to small fluctuations in the factors, that is, the results are the same for all experiments. Robustness testing is usually applied as the last test just before the release of a product or a method. When performing a robustness test of a method, the objective is

- to ascertain that the method is robust to small fluctuations in the factor levels,
- and, if non-robustness is detected…
- to understand how to alter the bounds of the factors so that robustness may still be claimed.

Robustness is achieved when the designer understands these potential sources of variation and takes steps to desensitize the product to them. G. Taguchi, a Japanese engineer, had a big effect on quality control and experimental design in the 1980s and 1990s. **The Taguchi Methods** (see http://www.stat.rutgers.edu/~buyske/591/lect10.pdf for more info.) is a well known strategy in robustness testing. Usually a fractional factorial (see section 2.2.3 in this editorial) or Plackett-Burman design is used.

3. Analysing the resulting experimental data

After the planning stage, when the set of experiments are laid out according to a statistical design, the planned experiments are made, either in parallel, or one after another. Each experiment gives results, i.e. values of the response variables. Thereafter, these data are analysed by means of multiple regression, or generalisations thereof such as the PLS and O-PLS methods (see earlier Editorials 2002). This gives a model relating the factors to the results, showing which factors are important, and how they combine in influencing the results. The model is then used to make predictions, e.g. how to set the factors to achieve desired (optimal) results.The fitted model is reviewed by…

- Examining the coefficients and their 95% confidence interval, or normal probability plots of effects and interactions.
- Examining the ANOVA table, check for curvature
- Plotting residuals, normal probability plot of residuals, and run order residuals
- Checking for the optimal transformation of the response through the use of the Box Cox plot.
- Conclusions:
  1. Select dominating factors
  2. Check and modify ranges
  3. Look for curvature

3.1 Multiple Linear Regression (MLR)

Traditionally, the most frequently used method for finding the regression coefficients **b** is the ordinary least squares method where:

$$\mathbf{y}=\mathbf{Xb}+\mathbf{f}$$
$$\mathbf{b}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

This minimizes the residuals ($\mathbf{f}^T\mathbf{f}$), which is equivalent to maximizing the fit to **y**. With MLR, the coefficients of the model are computed to minimize the sum of the squares of the residuals. In order to estimate **b**, MLR requires that the X-variables must be linearly independent ($(\mathbf{X}^T\mathbf{X})$ of full rank). It is important to also note that MLR fits one response at a time and hence assumes them to be independent.

3.2 Partial Least Squares Projections to Latent Structures (PLS)

PLS is one of the most common methods for analyzing multivariate data where a quantitative relationship between a descriptor matrix **X** and a response matrix **Y** is sought. The PLS model can be expressed by:

Model of **X**: $\mathbf{X} = \mathbf{TP}^T+\mathbf{E}$
Model of **Y**: $\mathbf{Y} = \mathbf{TC}^t+\mathbf{F}$

PLS contains MLR as a special case, then the PLS regression coefficients and the MLR coefficients are identical.

3.3 Orthogonal projections to latent structures (O-PLS)

The recent O-PLS methods [O-PLS and O2-PLS] (see Editorial from April 2002 on O-PLS), are improved modifications of the NIPALS PLS algorithm. The development of O-PLS has, like the orthogonal signal correction (OSC) filters (March 2002 Editorial), been driven by the large amount of non-correlated variation present in the data sets today, especially in a multivariate calibration situation. The interpretational ability of the other inverse regression models (PLS, PCR, MLR) largely depends on the degree of systematic orthogonal variation in **X** with regards to **Y**.

The basic idea of O2-PLS is to divide the systematic part in **X** and **Y** into two parts, one which is related to **X** and **Y**, and one that is not. For each matrix, the latter is computed in a way that makes it orthogonal to the other matrix, i.e. completely independent. If **X** or **Y** contains strong but irrelevant variation, O2-PLS improves the interpretational ability of the parameters in the model, e.g. score plots, loading plots compared to the MLR and PLS methods (and other methods with similar properties such as ridge regression).
Thus the O2-PLS model can be written as a factor analysis model, where some factors (**T**) are common to both **X** and **Y**;

X model: $\mathbf{X} = \mathbf{TW}^T + \mathbf{T}_{\text{Y-ortho}}\mathbf{P}_{\text{Y-ortho}}{}^T + \mathbf{E}$
Y model: $\mathbf{Y} = \mathbf{UC}^T + \mathbf{U}_{\text{X-ortho}}\mathbf{P}_{\text{X-ortho}}{}^T + \mathbf{F}$
Prediction of **Y**: $\mathbf{Y}_{hat} = \mathbf{TC}^T$

3.4 ANOVA - Analysis of Variance

ANOVA breaks up sums of squares in components and compares their size with F-test.

SS_Y = SS_Regression + SS_Residual
SS_Residual =
SS_lack_of_fit + SS_pure_error

If center points exist, lack of fit indicates curvature. Lack of fit compares pure error (from replicated experiments) with residual error (model error). Residual plots indicate

- Outliers
- Curvature
- Need for transformations

4. Applications of DoE

Chemical synthesis

1. Synthetic steps
2. Work up and separation
3. Reagents, solvents, catalysts
4. Structure ' reactivity and properties.

Biotech industry

1. Pharmaceutics, formulation for drug delivery
2. Media development & optimization
3. Biochemistry, drug design
4. Analytical biochemistry, separation (HPLC,…), assay development and optimization
5. Pharmacology
6. Process optimization and control, fermentation, separation, purification

Cosmetic industry

1. Processes, production, separation, cleaning,…
2. Formulations, shampoos, nail polish, creams, perfumes, soaps, powders, …
3. Molecular structure, high potency, low toxicity, allergenicity

Drug industry

1. Pharmaceutics, formulation for drug release, hardness of pills,…
2. Organic chemistry, synthesis, drug design, …
3. Analytical chemistry, Separation [HPLC, …], resolution, speed.
4. Pharmacology
5. Process optimization and control, synthesis, fermentation, separations, …

Process industry

1. Process optimization and control (yield, purity, through put time, pollution, energy consumption)
2. Product quality and performance (material strength, warp, color, taste, odour)
3. Product stability versus process variation

5. References

1. G.E.P Box, W.G. Hunter and J.S. Hunter "Statistics for experimenters", John Wiley and Sons, Inc,., New York (1978)
2. G.E.P Box, N.R. Draper "Empirical model-building and Response surfaces", John Wiley and Sons, Inc,., New York (1987)
3. C.K. Bayne and I.B. Rubin "Practical Experimental Designs and optimizations methods for chemists", VCH Publishers, Inc., Deerfield Beach, Florida (1986)
4. Morgan E. - Chemometrics: Experimental Design. John Wiley & Sons, Inc., New York
5. Engineering Statistics Handbook (NIST) [http://www.itl.nist.gov/div898/handbook/index.htm] July 2002.
6. Carlson R. - Design and Optimization in Organic Synthesis. Elsevier science publishers, Amsterdam
7. Slide notes from Svante Wold (personal communication)
8. Design of Experiments: Principles and Applications, Umetrics Academy - L. Eriksson, E. Johansson, N. Kettaneh-Wold, C. Wikström, and S. Wold, ISBN 91-973730-0-1
9. Response Surface Methodology: Process and Product Optimization Using Designed Experiments, 2nd Edition Raymond H. Myers, Douglas C. Montgomery ISBN: 0-471-41255-

Further Reading

- Introduction to Design of Experiments -See http://www.umetrics.com/pdfs/books/DOEBook.pdf
- Introduction to DoE -See http://kingkong.me.berkeley.edu/html/pres_assets/pdfs/fracfact2.pdf
- DoE overview Engineering Statistics Handbook (NIST) -See http://www.itl.nist.gov/div898/handbook/pri/pri_d.htm
- Optimization designs, NIST http://www.itl.nist.gov/div898/education/dex/optdesgn/optdesgn.pdf
- How to Select Design of Experiments Software -See http://www.qualitydigest.com/nov98/html/doe.html
- Robust Design and Taguchi Method -See http://www.stat.rutgers.edu/~buyske/591/lect10.pdf

Appendix

Model Parameters

Goodness of fit statistics, information about model adequacy

PRESS, Predicted Residual Sum of Squares: Sum of squared differences between predicted and observed y-values (over all rounds)

- $Q2 = 1 - \frac{PRESS}{SS\_tot}$ Predictive power of the model (according to cross validation).

Q2 underestimates the goodness of fit. Q2 > 0.5 good, Q2 > 0.9 excellent

- $R2 = (1 - \frac{SS\_residual}{SS\_tot})$ Percent of the variation of the response explained by the model.

R2 overestimates the goodness of fit.

- $R2adj = 1 - \frac{MS\_res}{MS\_tot}$,

$$MS\_tot = \frac{SS\_tot}{N - c},$$ c = 1 (if model incl. constant.) else c = 0.

$$MS\_res = \frac{SS\_res}{N - p}$$

Dealing with and quantifying variability

How to use a distribution ?

1. Hypothesis testing and Significance testing
   only reveal similar/dissimilar samples. Not very useful
2. Confidence intervals, estimation
   How similar are the samples →Much more useful

1.) Hypothesis testing

Assumption: Distribution normal, mean μ, std. error s_e.

Null Hypothesis:

The process is believed to operate at an impurity level with $\mu = 2.0$

Data: 6 samples were used to estimate

m = 4.4. s_e=0.3

$$s_e = \frac{std.error}{\sqrt{N}} = 0.75 / \sqrt{6} = 0.3$$

Question: Do we reject of not the Null Hypothesis?

Using the data:

Compute the probability to find an average impurity level (m) of 4.4 on any given day, if true mean = 2.

$$t = \frac{(m - \mu)}{s_e} = (4.4 - 2.0) / 0.30 = 8.0$$

At 5 degree of freedom (N-1), this t-value corresponds to a probability level of about 0.0002. There is about 2 chances in ten thousand to observe a value of 4.4 if the mean really is 2.0

→ Reject the Null hypothesis

2.) Confidence interval estimation (much more useful)

$Z = \frac{(4.4 - \mu)}{0.30}$ is t-distributed with 5 degrees of freedom (N -1)

$t(df, \alpha) = t(5, 0.025) = 2.57$

The 95 % confidence interval for μ is:

$$m \pm t(df, \alpha) * s_e = 4.4 \pm (2.57 * 0.30) = 4.4 \pm 0.77$$

$3.6 \le \mu \le 5.2$ Much more useful

Previous Editorials on HoC

Check them out at HoC's website
http:/www.acc.umu.se/~tnkjtg/Chemometrics/Editorial

Stay alert for upcoming editorials on HoC

July 2002	No Editorial this month
June 2002	A statistician's view of the single-y PLS problem Dick Kleinknecht
May 2002	Wavelets in Chemometrics - compression, denoising and feature extraction Johan Trygg
April 2002	How to create an OSC filter for PLS and end up with a new generic modelling method, O-PLS Johan Trygg
March 2002	Everything you need to know about Orthogonal Signal Correction (OSC) filters - and how they can improve interpretation of your data Johan Trygg
Feb 2002	Have you ever wondered why PLS sometimes needs more than one component for a single-y vector? Johan Trygg

Do you want to write an Editorial on HoC? Email us, johan.trygg@chem.umu.se