Statistical Models and Multi-Criteria Optimization Techniques in Chromatography

Pierre **LEBRUN**[1], Bruno **BOULANGER**[2], Philippe **HUBERT**[1] **,** Benjamin **DEBRUS**[1]et Bernadette **GOVAERTS**[3]

1. Institut de Pharmacie/Analyse des médicaments, Université de Liège, BAT. B36 Chimie analytique, Avenue de L'hôpital,1. B4000 Liège, Belgique. plebrun@ulg.ac.be
2. Eli Lilly S.A., statistical and mathematical sciences, 1348 Louvain-la-Neuve
3. Institut de Statistique, *Université catholique de Louvain-la-Neuve*, Voie du Roman Pays, 20, B1348 Louvain-la-Neuve, Belgique

KEYWORDS: liquid chromatography, design of experiments, linear models, multi-criteria optimization, error propagation.

1. Introduction

In analytical chemistry, the chromatographic techniques are widely used in different fields of activity such as chemical, pharmaceutical, biomedical, environmental and food analysis. Thus, the selection of the most appropriate experimental conditions allowing the separation of compounds of interest in various matrices is a matter of a very particular interest. Pharmaceutical industries are of course concerned by these problems and are more especially interested by all new issues allowing to separate their compounds properly and quickly in order to quantify them. Indeed, this analytical step is a crucial phase during the development of new drugs.

Amongst the chromatographic techniques, the High Performance Liquid Chromatography (HPLC) is one of the most used techniques to fulfill this objective.

Actual developments of analytical methods in HPLC are often time consuming and not always under the perfect control of the analysts. This is mainly due to the number of parameters to manage to obtain acceptable separation conditions. The situation still becomes complicated when the matrix is complex and contains many compounds with physico-chemical properties which are not necessarily known. However, there are scientific evidences that this process can be accelerated and analysed in order to improve the knowledge on this domain.

The present work presented in this report is part of a project called **ADAM**, which stands for *Automated Development of Analytical Methods*. This acronym summarizes the fact that the desired result is to get a faster development in a more accurate way, with an automated procedure. The total time for the development of a new method should not exceed one night.

2. Objectives

The global objective of this report is to present a methodology allowing obtaining the best *tuning parameters* of a HPLC in order to get the best chromatogram in the range of the analytical conditions for a domain of possible mixtures. This global target can be subdivided into several smaller objectives.

The first objective aims at adequately model the retention times of peaks, i.e. maximizing the quality of the fit, avoiding correlation in the responses and residuals while minimizing the risk of overfitting. This should ensure to have good predictive quality for the models.

The second objective is to optimize chromatographic analytical conditions (resolution, retention times, peaks width, asymmetry and separation). This is achieved by multi-criteria optimization using desirability functions of Derringer.

The third objective is to investigate how predictive errors of the models propagate into Derringer desirability index using Monte-Carlo methods. The distribution of Derringer desirability gives clues about the quality of the predicted optimal conditions.

3. Methodology

3.1. Notation

We note $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_K)$ the $(N \times K)$ matrix of factors. Each $\mathbf{x}_k$ represents an impactfull tuning parameter of a HPLC. There is one factor setting $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{iK})$ for each chromatogram $(1 \le i \le N)$. Each factor $\mathbf{x}_k$ can be either continuous or discrete, and is characterized by n_k levels. Continuous factors are defined over a domain of interest $[L_k, U_k]$. The number of levels of $\mathbf{x}_k$ depends of the complexity of its effects on the response $\mathbf{Y}^{(p)}$. We note the experimental domain, representing all possible assignations of $\mathbf{X}$, as χ.

A full factorial design [1] has been applied on K=4 factors:

-$\mathbf{x}_1$: pH, continuous, 4 levels : 2,5,7,10
-$\mathbf{x}_2$: gradient time, continuous, 3 levels : 10, 20, 30 min.
-$\mathbf{x}_3$: column, discrete, 5 levels
-$\mathbf{x}_4$: solvent, discrete, 2 levels

A common practice in design of experiments is to code the factors $\mathbf{x}_k$ into the interval $[-1,1]$. It allows good interpretation of regression coefficients while removing the units of the factors.

Each observation is a chromatogram. A chromatogram is a kind of temporal curve which contains peaks at certain moments. One peak is the response of one compound. If a solution contains M compounds, then the corresponding chromatogram possesses M peaks. The vector $\mathbf{c} = (c_1, ..., c_j, ..., c_M)$ references compounds, i.e. peaks, in a chromatogram $(1 \le j \le M)$. Each peak is referenced with three positions: $\mathbf{B}$ (beginning) $\mathbf{A}$ (apex) and $\mathbf{E}$ (end). $\mathbf{B}$, $\mathbf{A}$ and $\mathbf{E}$ are vectors containing the original responses to be modeled. Their values are the observed time at the beginning, apex and end of each peak, for each chromatogram. Let's define the vector containing all the beginnings of peak $\mathbf{B}$ as

$$\mathbf{B} = ((B_{11}, ..., B_{1j}, ..., B_{1M}), ..., (B_{i1}, ..., B_{ij}, ..., B_{iM}), ..., (B_{N1}, ..., B_{Nj}, ..., B_{NM})).$$

$\mathbf{A}$ and $\mathbf{E}$ are defined in the same way.

3.2. Responses

High correlations between original responses $\mathbf{B}$, $\mathbf{A}$ and $\mathbf{E}$ exists. It is surely a good idea to find other responses to model. Moreover, when predicting responses, uncertainty can lead to inversion between beginning, apex and end of peaks, which is not desirable. Let us define the P responses to model as transformation of the original ones $(1 \leq p \leq P)$. Back transformation must be possible.

$$\mathbf{Y}^{(p)} = f_p(\mathbf{B},\mathbf{A},\mathbf{E})$$
$$(\mathbf{B},\mathbf{A},\mathbf{E}) = f^{-1}(\mathbf{Y}^{(1)},\mathbf{Y}^{(2)},...,\mathbf{Y}^{(p)})$$

3.3. Models

We want to model the transformed responses $\mathbf{Y}^{(p)}$ against the chosen relevant factors (tuning parameters of HPLC). We decide to apply functions like

$$\mathbf{Y}_{ij}^{(p)} = g_p(\mathbf{x}_i, c_j; \boldsymbol{\beta}_p) + \varepsilon_{ij}^{(p)}.$$

Typically, the function g_p can be written as a multiple linear regression [8].

3.4. Optimum finding

In classical optimization problems, the target is often to maximize or minimize the responses of the models. In our case, we want to maximize several criteria computed from predicted values of $\hat{\mathbf{B}}$, $\hat{\mathbf{A}}$ and $\hat{\mathbf{E}}$. The different criteria $\mathbf{cr}_z$ (1<z<Z),to be simultaneously optimized, are thus computed from t_z functions as

$$\mathbf{cr}_z = t_z\left(\mathbf{B}_j, \mathbf{A}_j, \mathbf{E}_j; j = 1,...,M\right)$$

Only one value for each criteria is kept for each chromatogram. We chose well-known criteria in chromatography: minimal resolution, minimal separation, maximal width of peaks, maximal retention times and maximal asymmetry.

To perform a multi-criteria optimization process, we use desirability functions of Derringer [3][5], combined with a geometric mean, giving the global desirability index:

$$D(\mathbf{cr}) = \left(\prod_{z=1}^{Z} d_z(cr_z)^{w_z}\right)^{1/Z} \quad \text{with} \sum_{z=1}^{Z} w_z = 1.$$

d_z are functions based on the Cumulative Distribution Function (CDF) of a Normal distribution [6][7]. Characterization of the Normal (means, standard deviations) must be adequately chosen, on the basis of the distribution of the criteria across the experimental domain. d_z transforms the criteria cr_z into its desirability form. d_z (cr_z) is equals to one is the criteria is perfectly fulfilled, and to zero if the criteria is the worst possible solution.

We want to find the factor setting **x*** that maximize the value of D(cr). We chose to use the grid-search method [9], consisting in trying all possibilities of factor settings ($\mathbf{x}_0$) belonging to the experimental domain.

$$\mathbf{x}^* = \left(\max_{\mathbf{x}_0 \in \chi} \prod_{z=1}^{Z} (d_z(t_z(f^{-1}(g_1(\mathbf{x}_0, \mathbf{c}, \hat{\boldsymbol{\beta}}_1),...,g_p(\mathbf{x}_0, \mathbf{c}, \hat{\boldsymbol{\beta}}_p)))))^{w_z}\right)^{1/Z}$$

The Figure 1 shows an optimized chromatogram. Optimal factors are a pH of 6, a gradient time of 28 minutes, a Xbridge C8 column and CH_3CN as solvent in the buffer.

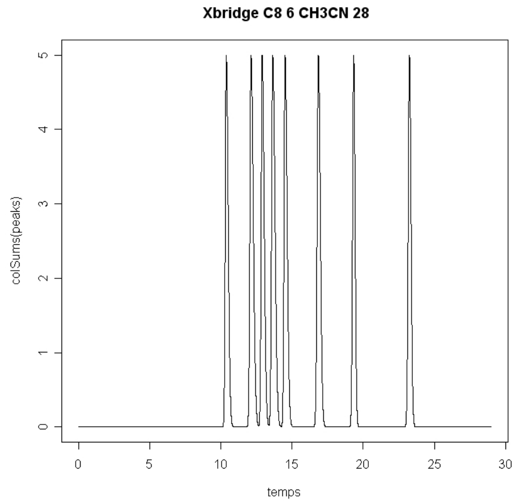

Figure 1. Multi-criteria optimized chromatogram.

3.5. Error propagation

To assess the quality of an optimal solution, one can generate an error around it, using Monte-Carlo methods. We propagate the mean response predictive error of the models across the transformation of responses, across the criteria, to the global desirability index. Firstly, we generate a random normal error on prediction around optimal solution.

$$\hat{Y}_{sim}^{(p)} = \hat{Y}^{*(p)} + N(0, \hat{\sigma}^2_{\hat{Y}^{*(p)}})$$

We can then propagate the simulated errors across original responses, across the criteria to the global desirability index.

$$\begin{aligned}
(\hat{\mathbf{B}}_{sim}, \hat{\mathbf{A}}_{sim}, \hat{\mathbf{E}}_{sim}) &= f^{-1}(\hat{\mathbf{Y}}_{sim}^{(1)}, \ldots, \hat{\mathbf{Y}}_{sim}^{(P)}) \\
\mathbf{c}\hat{\mathbf{r}}_{z,sim} &= t_z(\hat{\mathbf{B}}_{sim}, \hat{\mathbf{A}}_{sim}, \hat{\mathbf{E}}_{sim}), \quad z = 1, \ldots, Z \\
D(\mathbf{c}\hat{\mathbf{r}})_{sim} &= \left(\prod_{z=1}^{Z} d_z(\mathbf{c}\hat{\mathbf{r}}_{z,sim})^{w_z} \right)^{1/Z}
\end{aligned}$$

Distribution of criteria can give good visual idea of the form taken by the error. The figure 2 presents the distribution of criteria around an optimal solution. The form and the position of the distributions can give clues about the optimal solution.

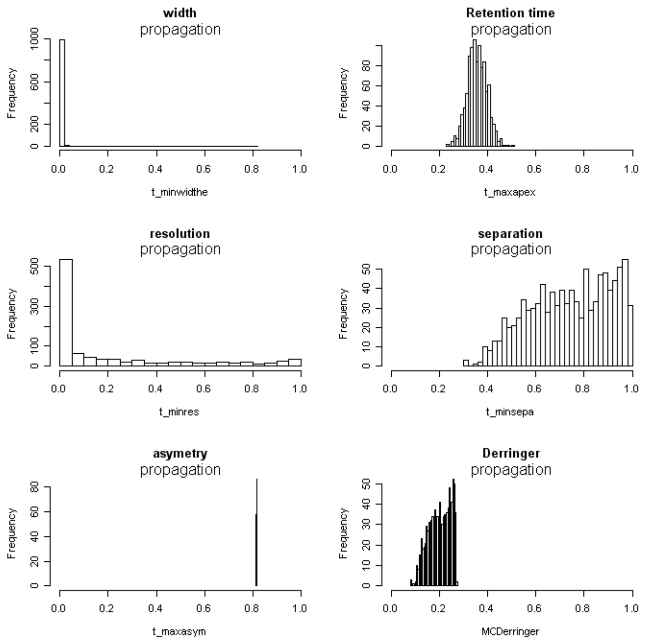

Figure 2. **Distribution of the criteria under the error of the statistical models**

However we want to answer this question: considering the error (experimental error + model error), is the optimal solution really better than other suboptimal solutions? We proceed as follow:

1. Re-estimate D(cr) on the whole *quantitative* experimental domain using grid method. For this, qualitative factors are fixed to optimum. A contour plot of the global desirability can be realized if there is 2 quantitative factors.
2. Locate local and global optimal conditions.
3. Simulate data with Monte-Carlo around the global optimum **x*** ; compute the criteria and the global desirability $D(\mathrm{cr})_{\mathrm{sim}}$
4. Find the 5th percentile of $D(\mathrm{cr})_{\mathrm{sim}}$.
5. Differentiate contour lines on the contour plot for values greater than the 5th percentile of $D(\mathrm{cr})_{\mathrm{sim}}$.

Then, the objective is to see if the contour lines added in the 4th step are taking some local positions in the overall desirability. Removing the 5 first percentiles of the distribution allows omitting some outlier values which can destabilize results and give a false opinion. This is presented in figure 3.

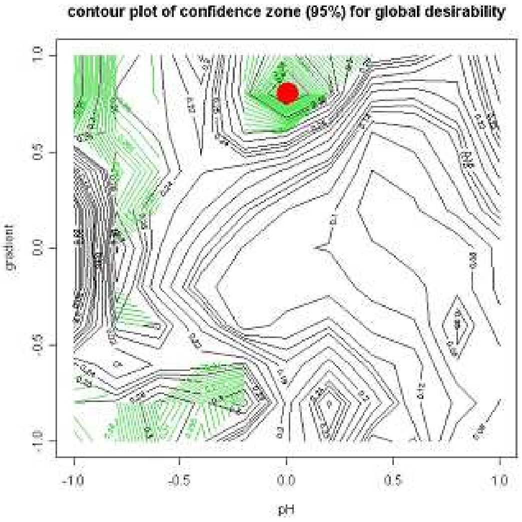

Figure 3. Illustration of the presented methodology to assess the quality of a predicted optimal solution (red point)

4. Case study

The presented methodology is applied on real data. Models are not easy to be fitted due to experimental error still to be estimated, and errors in the data file can occur. The most problematic error consists in inversing the identification of two peaks in one chromatogram.

Experiences are performed by Benjamin Debrus [2], from University of Liège, in collaboration with Eli Lilly S.A. Experiences consists in trying various analytical conditions described by the design of experiments. Chromatogram and identified peaks are obtained from these experiments. Results on models and optimization process will be presented.

Bibliography

[1] G.M. Cox and W. Cochran (1957). *Experimental Designs*,2nd Edition. Wiley.

[2] B. Debrus(2006), *Contribution au développement automatisé de methods analytiques*. Master's thesis, University of Liege.

[3] G.C. Derringer and R. Suich (1980). *Simultaneous Optimization of Several Response Variables*. J.Qual.Tech.,12(4):214–219.

[4] W.Dewe and al.(2004), *Development of Response Models for Optimizing HPLC Methods*. Elsevier, Chemometrics and Intelligent Laboratory Systems 74:263–268.

[5] E.C.Harrington (1965).*The desirability function*. Industrial Quality Control, 21:494–498.

[6] C.Le Bailly de Tilleghem and B.Govaerts (2005). *Distribution of Desirability Index in Multi-Criteria Optimization using Desirability Functions based on the Cumulative Distribution Function of the Standard Normal*. Institute of Statistics, UCL.

[7] C. Le Bailly de Tilleghem and B. Govaerts (2006). *Uncertainty Propagation in Multiresponse Optimization using a Desirability Index*. Institute of Statistics, UCL.

[8] J. Neter, W. Wasserman, and M.H. Kutner (1990). *Applied Linear Statistical Models,* third edition. Irwin.

[9] S.J. Russell and P. Norvig (2002). Artificial Intelligence: A Modern Approach (2ndEdition). PrenticeHall.