Use of Bayesian multivariate prediction models to optimize chromatographic methods

Pierre Lebrun, ULg & Arlenda
Bruno Boulanger, ULg & Arlenda
Philippe Lambert, ULg & UCL
Astrid Jullion, UCB Pharma

UCB Pharma
Braine l'Alleud (Belgium)
May 2010

Overview

- ICH Q8 regulatory document
  - Design Space definition
  - Risk based approach
- Classical optimisation approach
  - Drawbacks
- Bayesian approach
  - Predictive distribution
  - Using informative prior distributions
- Example
  - Optimization of a chromatographic method
    - Model
    - Predictive distribution under informative prior distribution of parameters
    - Monte-Carlo simulations for multi-criteria decision method
- Conclusions

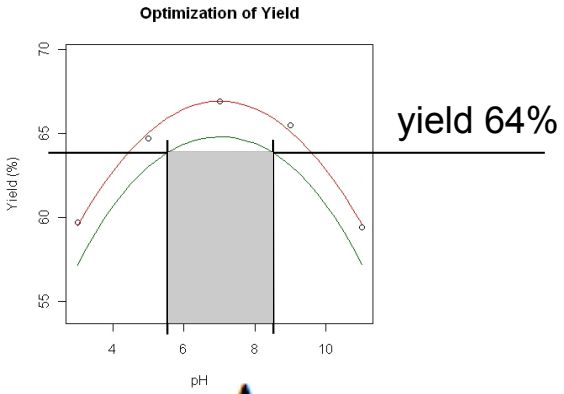
ICH Q8

- Target : *Understand and gain knowledge about a process/method to find a parametric region of reliable robustness for future performance of this process/ method* -> **assurance of quality**

- This region is the Design Space

$$\mathrm{DS} = \{\mathbf{x}_0 \in \chi \mid E_{\theta,data}\left[P(\mathbf{Y}(\mathbf{x}_0) \in \mathbf{\Lambda}) \mid \mathbf{x}_0, \boldsymbol{\theta}\right] \geq \pi\}$$

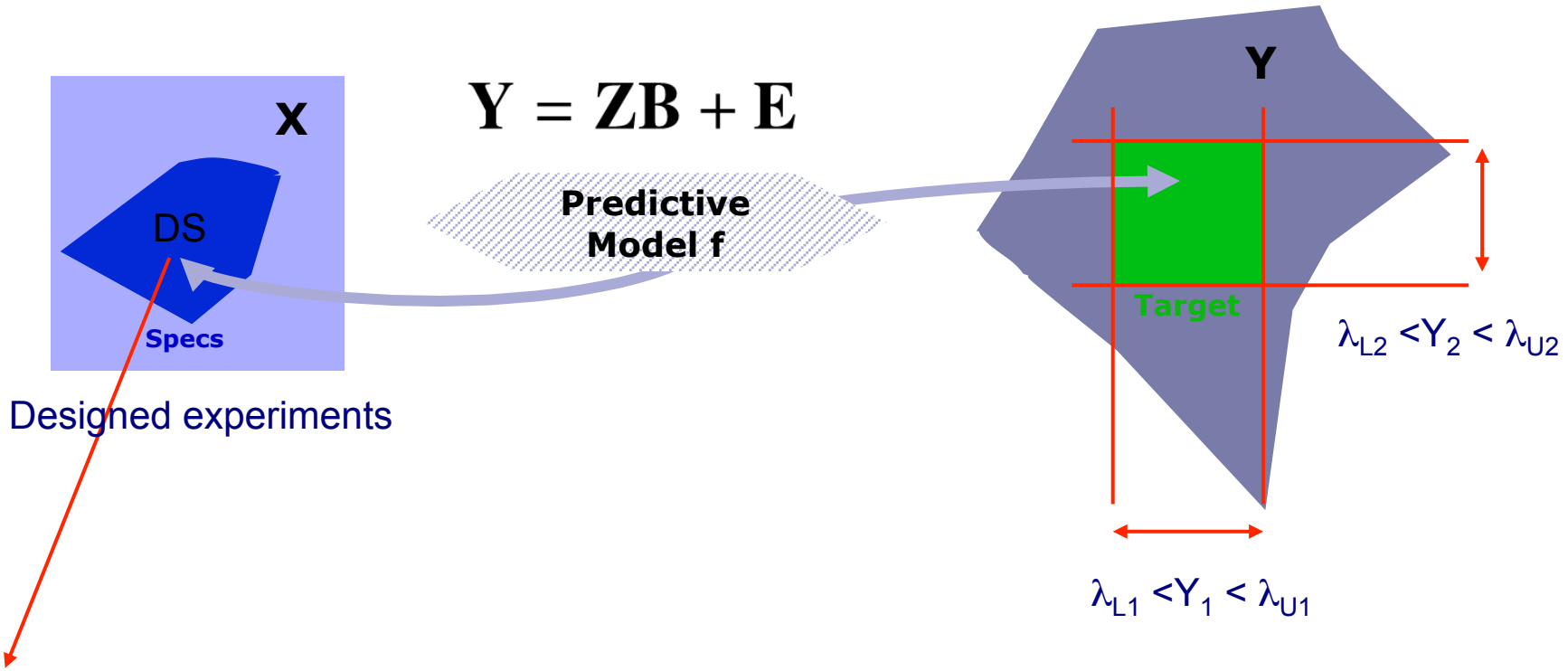
χ : domain
$\boldsymbol{x}_k$: set of combinations of process parameters
$\boldsymbol{Y}(\boldsymbol{x}_k)$: responses obtained for the $\boldsymbol{x}_k$ condition
$\boldsymbol{\Lambda}$: pre-defined set of acceptance limits
π_{min} : quality level (min. probability to achieve $\boldsymbol{\Lambda}$)



- We are also interested in the risk not achieving $\boldsymbol{\Lambda}$

ICH Q8

- Application (Boulanger et al., NCB09, Boston)



$$DS = \{\mathbf{x}_0 \in \chi \mid E_{\theta,data}\,[P(\mathbf{Y}(\mathbf{x}_0) \in \Lambda) \mid \mathbf{x}_0, \theta] \geq \pi\}$$

Classical (optimisation) approach

- In applied DoE literature, it is frequent to see the term "Design space"
  - (as the design of experiment itself…)
  - as the zone where **mean responses** satisfy acceptance limits Λ

  → But, mean responses
    → **do not provide any clue** about process reliability
    → **fail to give any information** on how the process will perform in the **future**
    → will certainly **give disappointing and unexplained results** for the future use of the process/method !

  ➔ ICH Q8 definition of DS is not met

Friends don't let friends use "overlapping means" to calibrate an ICH Q8 design space, J. Peterson, NCB 09, Boston

Classical optimisation approach

- Curse of dimensionality
  - Using classical (frequentist) multivariate models
    - Many responses (M) and many parameters (F)
    - Cost of experiments leads to *light* DoE (low N)
    - *d.f.* : $\nu = N-(F+M)+1$ => possibly a negative value !
- (Predictive) Tolerance intervals

  "*In the theory of statistical tolerance regions, as usually presented in frequentist terms, there are inherent difficulties of formulation, development and interpretation*"
  Aitchison, Bayesian Tolerance Intervals, 1964

- A (posterior) predictive approach must be envisaged
  - Gain information through prior knowledge
  - Takes into account model and data uncertainty
  - Easier interpretation of results

Bayesian Design Space

- Bayesian analysis is well suited for
  - Standard multivariate regression,
  - Seemingly unrelated regression, non-linear, random effect, etc.
  - In **simple cases**, a predictive distribution of the responses can be identified and easily used
  - In **complex cases**, MCMC simulations from the likelihood and the parameter prior distributions are required
  - In **less complex cases**, sampling from identified parameters posterior distributions are used

- Bayesian computations
  - Posterior: $P(\theta|data) = P(data|\theta).P(\theta)/P(data) \propto P(data|\theta) \; . \; P(\theta)$

    $\propto$ Likelihood . Prior
  - Predictive: $P(Y^*|data) = \int P(Y^*|\theta) \; . \; P(\theta|data) \, d\theta$

Bayesian multivariate regression

- For the *M*-responses model (Box & Tiao 1973, Press 2003, Peterson 2004),

$$\mathbf{Y} = \mathbf{ZB} + \mathbf{E} \qquad \boldsymbol{\varepsilon}^{n'} \sim N_M(\mathbf{0}, \boldsymbol{\Sigma}), \quad n = 1, \ldots, N$$

- using *non-informative* prior distribution of the parameters,

$$p(\mathbf{B}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{1}{2}(M+1)}$$

- posterior distributions can be computed (Bayes theorem),

$$\mathbf{A} = (\mathbf{Y} - \mathbf{Z}\hat{\mathbf{B}})'.(\mathbf{Y} - \mathbf{Z}\hat{\mathbf{B}})$$

$$\boldsymbol{\Sigma} \mid \text{data} \sim W_M^{-1}(\mathbf{A}, \nu + M - 1), \quad \nu > 0$$

$$\mathbf{B} \mid \boldsymbol{\Sigma}, \text{data} \sim N_{F \times M}\left(\hat{\mathbf{B}}, \boldsymbol{\Sigma}, (\mathbf{Z}'\mathbf{Z})^{-1}\right)$$

$\nu = N-(F+M)+1$

OLS estimate of **B**

Bayesian multivariate regression

- The predictive distribution of responses at $\mathbf{x}_0$ is identified as a multivariate Student distribution:

$$p(\mathbf{Y}(\mathbf{x}_0) \mid data) = \iint p(\mathbf{Y} \mid \mathbf{x}_0, \mathbf{B}, \boldsymbol{\Sigma}) . p(\mathbf{B}, \boldsymbol{\Sigma} \mid data)\, d\mathbf{B} d\boldsymbol{\Sigma}$$

$$\mathbf{Y}(\mathbf{x}_0)|data \sim T_M\left(\hat{\mathbf{B}}\mathbf{z}_0, \left(1 + \mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0\right)\frac{\mathbf{A}}{\nu}, \nu\right)$$

$\frac{\mathbf{A}}{\nu}$ is the estimated covariance matrix

- Now, what if informative priors are used ?
  - Conjugate prior distributions :

$$\boldsymbol{\Sigma} \sim W_M^{-1}(\boldsymbol{\Omega}, \nu_0)$$

Prior scale matrix

Prior d.f. : $\nu_0 = N_0 - (M + F) + 1$

$$\mathbf{B} \mid \boldsymbol{\Sigma} \sim N_{(F \times M)}(\mathbf{B}_0, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}_0)$$

Prior mean parameters

Prior precision matrix of the parameters, (common for each response)

Bayesian multivariate regression - informative

- Posterior distribution can be identified (Bayes theorem),

$$\mathbf{B} \mid \boldsymbol{\Sigma}, \text{data} \sim N_{F\times M}\left(\mathbf{M}_{\mathbf{B}\text{post}}, \boldsymbol{\Sigma}, \left(\mathbf{Z}'\mathbf{Z} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}\right)$$

$$\mathbf{M}_{\mathbf{B}\text{post}} = \left(\mathbf{Z}'\mathbf{Z} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}\left(\mathbf{Z}'\mathbf{Z}\hat{\mathbf{B}} + \boldsymbol{\Sigma}_0^{-1}\mathbf{B}_0\right)$$

$$\boldsymbol{\Sigma} \mid \text{data} \sim W^{-1}(\boldsymbol{\Omega} + \mathbf{A}^*, \nu + N_0)$$

$$\mathbf{A}^* = \mathbf{Y}'\mathbf{Y} + \mathbf{B}_0'\boldsymbol{\Sigma}_0^{-1}\mathbf{B}_0 - (\mathbf{Z}'\mathbf{Z}\hat{\mathbf{B}} + \boldsymbol{\Sigma}_0^{-1}\mathbf{B}_0)'(\mathbf{Z}'\mathbf{Z} + \boldsymbol{\Sigma}_0^{-1})^{-1}(\mathbf{Z}'\mathbf{Z}\hat{\mathbf{B}} + \boldsymbol{\Sigma}_0^{-1}\mathbf{B}_0)$$

- With some (tedious) computations, it is possible to find the predictive distribution of a new response at $\mathbf{x}_0$:

$$\mathbf{Y}(\mathbf{x}_0)|data \sim T_M\left(\mathbf{M}_{\mathbf{B}\text{post}}\mathbf{z}_0, \left(1 + \mathbf{z}_0'(\mathbf{Z}'\mathbf{Z} + \boldsymbol{\Sigma}_0^{-1})^{-1}\mathbf{z}_0\right)\frac{\boldsymbol{\Omega} + \mathbf{A}^*}{\nu + N_0}, \nu + N_0\right)$$

Bayesian multivariate regression

- This predictive distribution is of particular interest as
  - There is **no need to simulate** from the prior or even the posterior distribution of parameters
    - No convergence issue in MCMC
  - It gathers the **uncertainty** of data and model parameters for a new responses vector
  - Quantiles of the multivariate Student are **β-expectation tolerance intervals** (Guttman, 1969)
  - It is a **generalisation** of the multivariate Student distribution for non-informative prior distributions

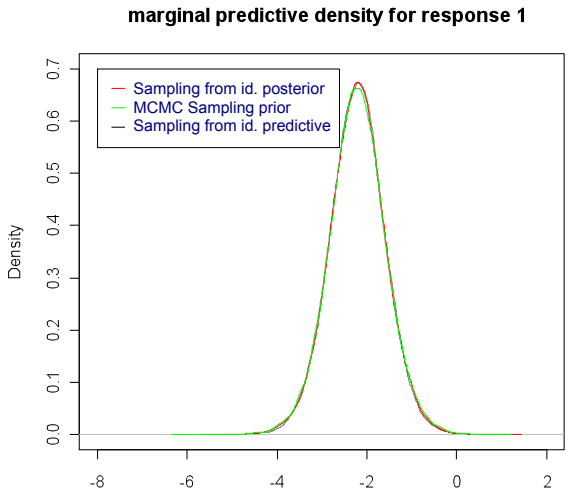
Verification – densities comparison

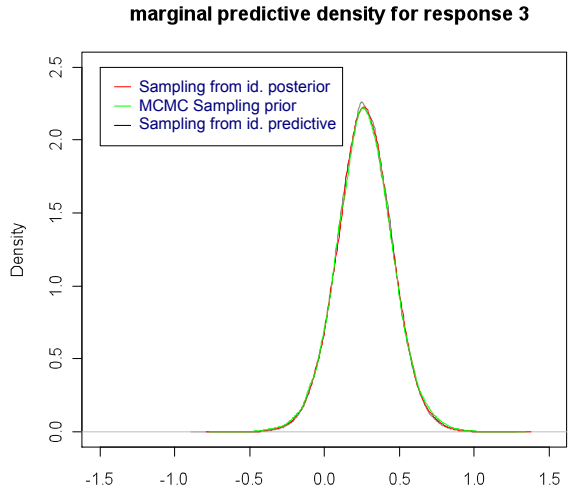
- Numerically get the posterior distribution using MCMC on prior distribution and likelihood (e.g. Winbugs)
- Direct sampling from the identified posterior distribution of parameters

$$\text{SAMPLE } (\mathbf{B}^s, \boldsymbol{\Sigma}^s) \text{ FROM } p(\mathbf{B}, \boldsymbol{\Sigma} \mid data)$$

Samples from the predictive distribution

- Direct sampling from the identified predictive distribution





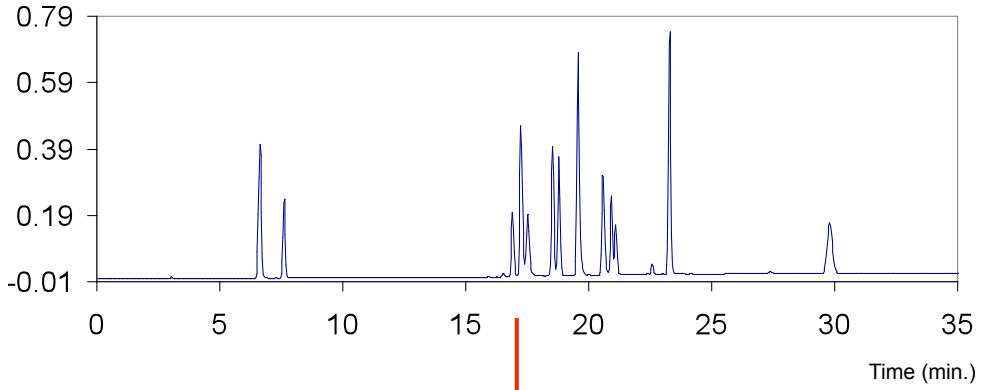
(Simulated data)
Densities are similar
whatever computation types !

Overview

- ICH Q8 regulatory document
  - Design Space definition
  - Risk based approach
- Classical optimisation approach
  - Drawbacks
- Bayesian approach
  - Predictive distribution
  - Using informative prior distributions
- Example
  - Optimization of a chromatographic method
    - Model
    - Predictive distribution under informative prior distribution of parameters
    - Monte-Carlo simulations for multi-criteria decision method
- Conclusions

Example

- Chromatographic method

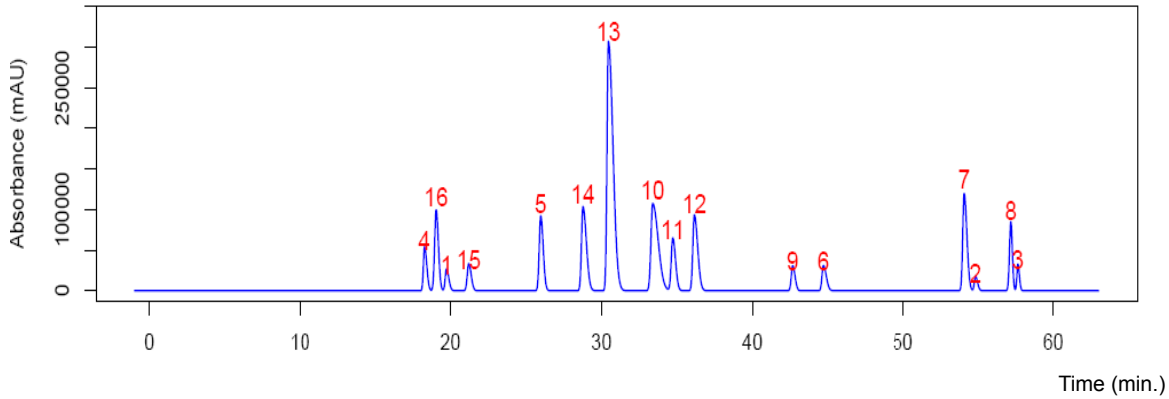


'Bad' chromatograms are difficult to interpret and to use

Peaks correspond to analytes

Tuning parameters of the system (HPLC)

Complex problem addressed with DoE



Good chromatograms with identified compound behind each peak
→ further analyses are possible

Example

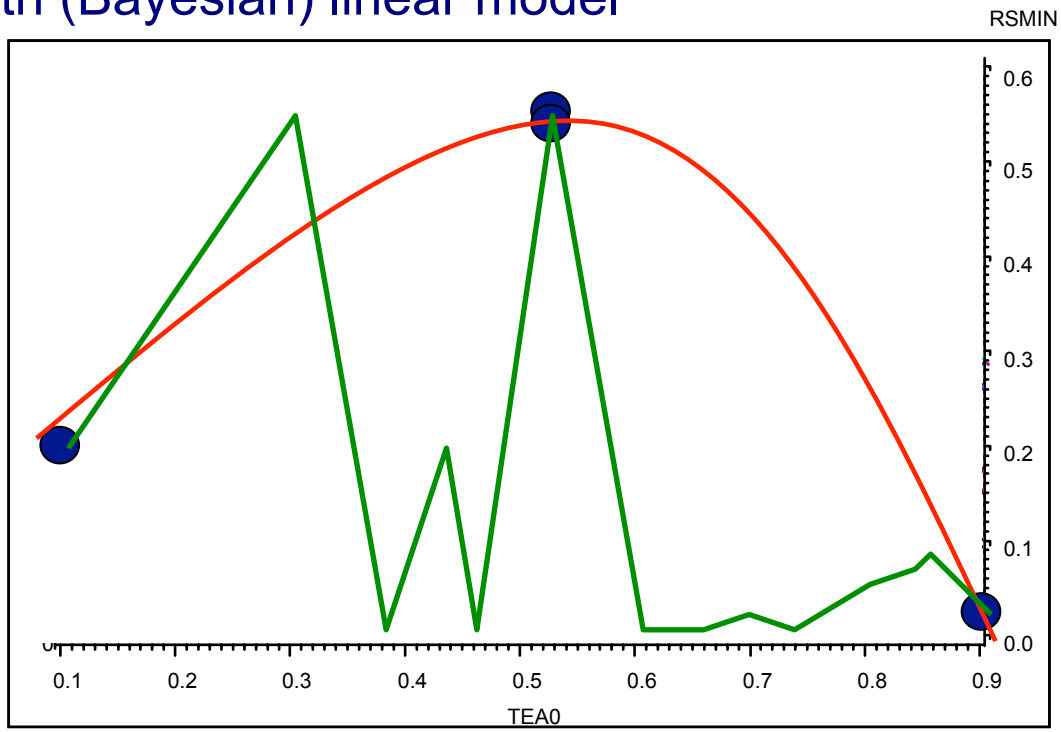
- We are interested to find the set of parameters settings that will give satisfactory results in the future use of this method (=DS).
  - ‘Satisfactory’ means ‘good’ chromatograms with well separated and nice-shaped peaks, and short run time, if possible

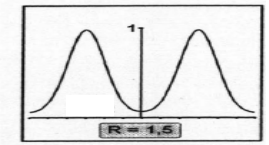
  - ➔ Potentially, **many responses** are modelled together
    - ➔ Each peak = 3 responses
    - ➔ Even **more responses than experiments** !

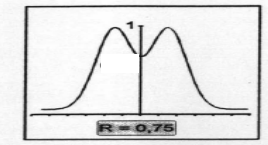
  - ➔ However, clear **correlation structure** exists among these responses
  - ➔ Linear relationship is assumed between (transformed) responses and predictors

(Counter)-example – Current practice

- The use of Bayesian methods is of no help if responses are not carefully chosen
  - Ex: quality criterion of interest (minimal resolution) is modelled with (Bayesian) linear model







Points of the DoE

What Existing Criteria ?

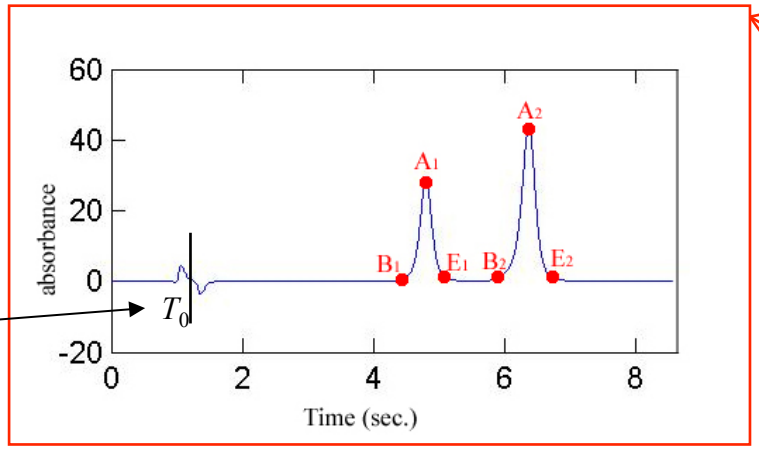
$$\min_{1 \le j \le P-1} \left(\frac{2 * |A_{(j+1)} - A_{(j)}|}{(E_{(j+1)} - B_{(j+1)}) + (E_{(j)} - B_{(j)})} \right)$$

Example – Responses choice

It is advised to model responses that show nice modelling properties (Massart et al. 1997, Snyder et al. 1997)

- Even if they are not directly related to 'quality'
- Quality criteria must be computable from the selected responses

Dead time of the chromatographic system



$$Y_{a1} = \log\left(\frac{A_1 - T_0}{T_0}\right)$$

$$Y_{b1} = \log\left(\frac{B_1 - T_0}{T_0}\right)$$

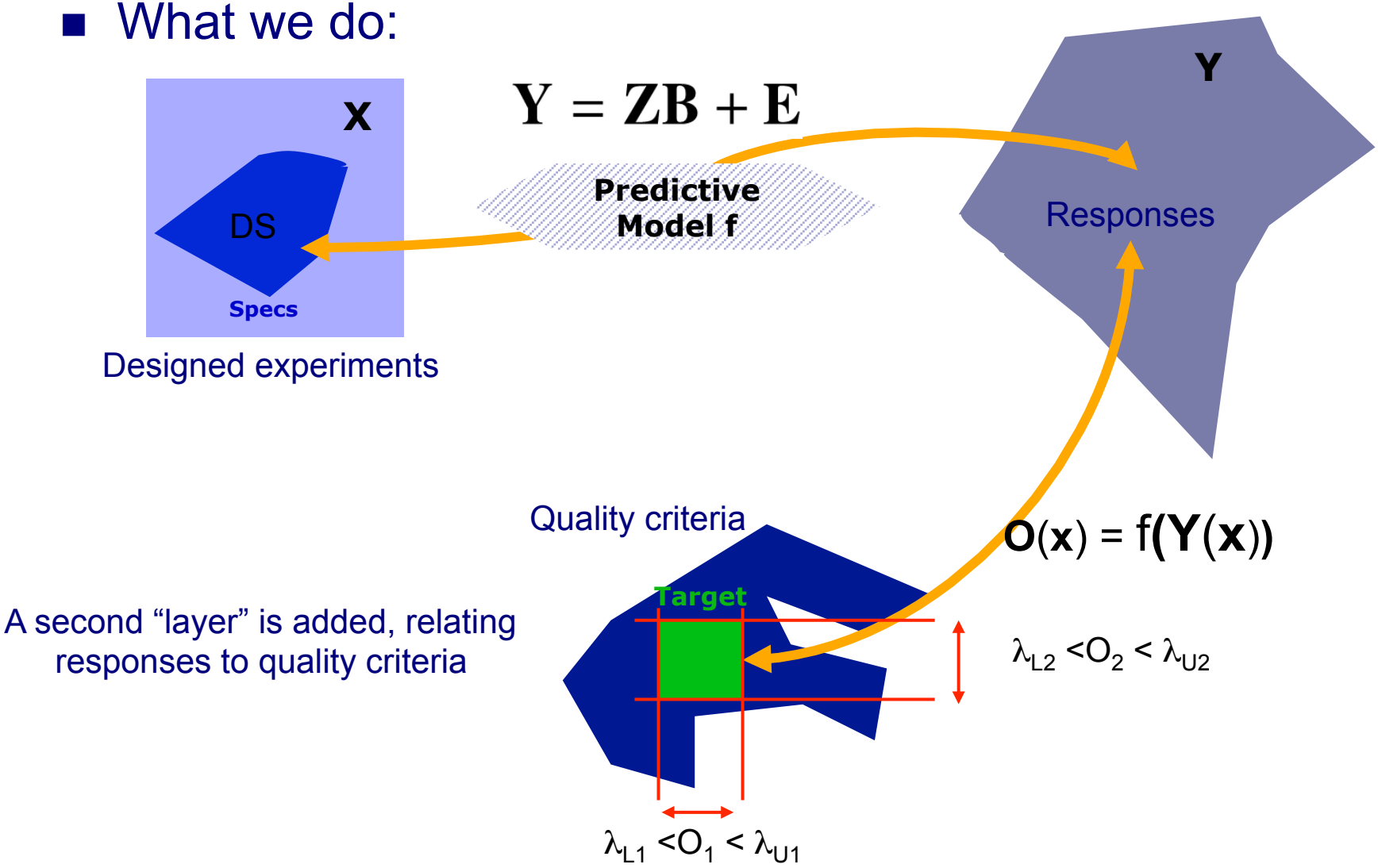
$$Y_{e1} = \log\left(\frac{E_1 - T_0}{T_0}\right)$$

$$\mathbf{Y} = g(B_1, A_1, E_1, \ldots, B_3, A_3, E_3)$$

We assume M responses (=3*P peaks) are observed $\underset{(N\text{x}M)}{\mathbf{Y}} = \underset{(N\text{x}F)(F\text{x}M)}{\mathbf{ZB}} + \mathbf{E}$

Example

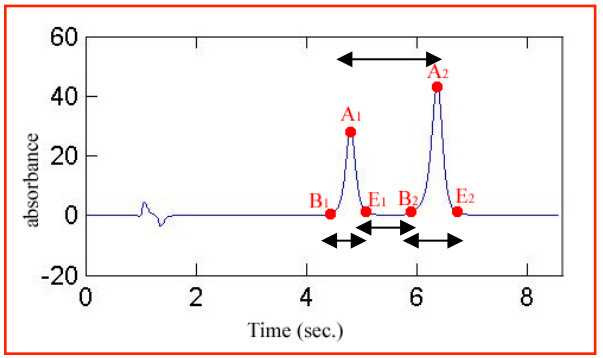
- What we do:



A second “layer” is added, relating responses to quality criteria

Example – Quality criteria

- Second layer
  - Combinations of responses



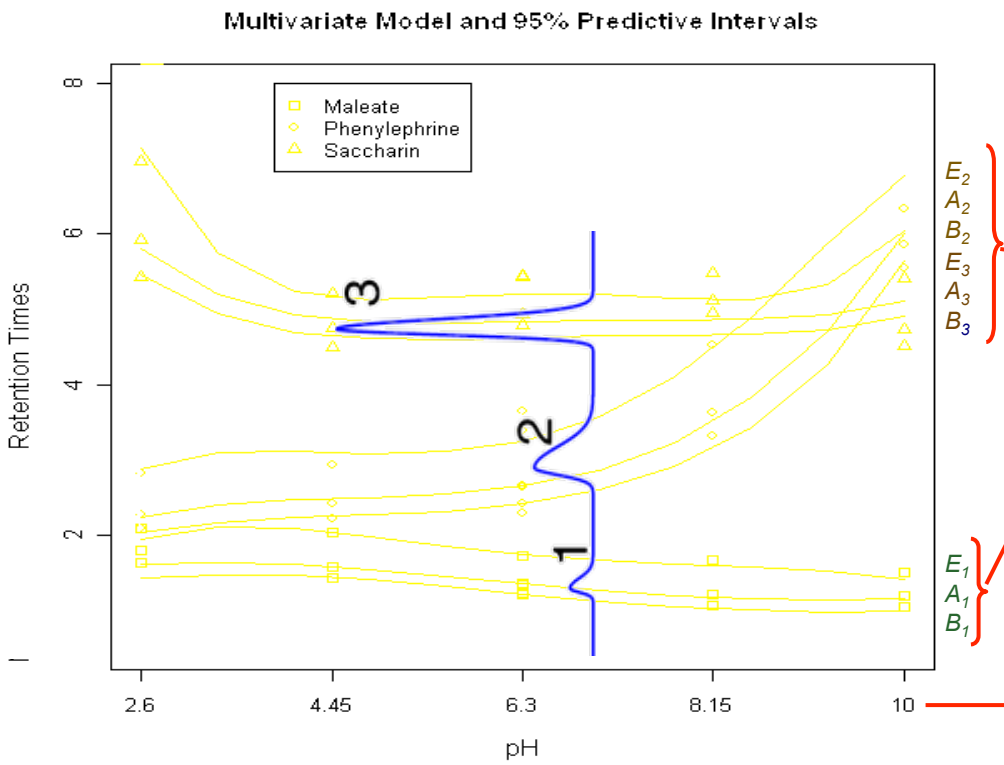
Ex:

$$\underbrace{\begin{vmatrix} O_1 = \text{Max. Time} & = \max\limits_{1\le j\le P}(A_j) & < \lambda_1 \\ O_2 = \text{Min. Separation} & = \min\limits_{1\le j\le P-1}(B_{(j+1)} - E_{(j)}) & < \lambda_2 \\ \dots & \dots & \dots \\ O_Z = \text{Min. Resolution} & = \min\limits_{1\le j\le P-1}\left(\frac{2 * |A_{(j+1)} - A_{(j)}|}{(E_{(j+1)} - B_{(j+1)}) + (E_{(j)} - B_{(j)})}\right) & < \lambda_Z \end{vmatrix}}_{\boldsymbol{O}} \quad \underbrace{\qquad}_{\boldsymbol{\Lambda}}$$

$$\text{DS} = \{\mathbf{x}_0 \in \chi \mid E_{\mathbf{B},\boldsymbol{\Sigma}|data}\left[P(\mathbf{O}(\mathbf{x}_0) \in \boldsymbol{\Lambda}) \mid \mathbf{x}_0, \mathbf{B}, \boldsymbol{\Sigma}\right] \ge \pi\}$$

→ DS is the set of conditions, such that the predictive probability that **O**bjectives will be simultaneously (jointly) within the acceptance limits is higher than π_{min}

Example – 3 peaks



M=9 responses are modelled with the Bayesian multivariate regression

A slice of the design is represented - DoE involves 2 factors :

- pH (quadric)
- Gradient time (quadratic)
+ interactions

Is there a zone in my domain where I can guarantee a satisfactory chromatogram in the future use of the chromatographic method ?

Prior distribution of B

- Model : set up of informative prior distributions
  - Assume no knowledge on the conditional distribution of $\mathbf{B}$

$$\mathbf{B} \mid \boldsymbol{\Sigma} \sim N_{(F \times M)}(\mathbf{B}_0, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}_0)$$

Matrix of 0 or $\hat{\mathbf{B}}$ are typical choices

A 'flat' diagonal (*F*x*F*) matrix (~**0** precision)

Prior distribution of Σ

- We know some correlations exist between responses (in Σ)
  - Correlations **within a peak: strong**; between peaks: none
  - However we don't have clue about covariance or scale

$$\boldsymbol{\Sigma} \sim W_M^{-1}(\boldsymbol{\Omega}, \nu_0)$$

ν_0 : degree of certainty of the prior
→ as low as possible (e.g. 3 or 4)

The desired correlation matrix is rescaled so it is comparable to the estimated scale matrix **A*≡** $\{a_{ij}^*\}$

$$\boldsymbol{\Omega}_{\text{cor}} = I_P \otimes \begin{bmatrix} \rho_1 & \rho_2 & \rho_2 \\ \rho_2 & \rho_1 & \rho_2 \\ \rho_2 & \rho_2 & \rho_1 \end{bmatrix} \Leftrightarrow \boldsymbol{\Omega}_{\text{cor}} = \begin{pmatrix} \begin{pmatrix} \rho_1 & \rho_2 & \rho_2 \\ \rho_2 & \rho_1 & \rho_2 \\ \rho_2 & \rho_2 & \rho_1 \end{pmatrix} & 0 & 0 \\ 0 & \begin{pmatrix} \rho_1 & \rho_2 & \rho_2 \\ \rho_2 & \rho_1 & \rho_2 \\ \rho_2 & \rho_2 & \rho_1 \end{pmatrix} & 0 \\ 0 & 0 & \begin{pmatrix} \rho_1 & \rho_2 & \rho_2 \\ \rho_2 & \rho_1 & \rho_2 \\ \rho_2 & \rho_2 & \rho_1 \end{pmatrix} \end{pmatrix}$$

E_1 A_1 B_1 E_2 A_2 B_2 E_3 A_3 B_3

Between peaks structure

Within peak structure

$$\begin{cases} \rho_1 = 1 \\ \rho_2 = 0.9 \end{cases}$$

$$\boldsymbol{\Omega} \equiv \{\Omega_{ij}\} = \frac{\sqrt{\{a_{ii}^*\}} \cdot \{\Omega_{\text{cor},ij}\} \cdot \sqrt{\{a_{jj}^*\}}}{\nu} * N_0$$

Prior distribution of Σ

- Thus, responses are correlated 3 by 3
  - Why not one model for each peak (3 responses only) ?
  - For the sake of generality, it is enviable to have a model that can handle any correlation structure
    - Two different but structurally very similar compounds will have responses that will be correlated (e.g. enantiomers)

$$\Omega_{cor} = \begin{pmatrix} \begin{pmatrix} \rho_1 & \rho_2 & \rho_2 \\ \rho_2 & \rho_1 & \rho_2 \\ \rho_2 & \rho_2 & \rho_1 \end{pmatrix} & 0 & \begin{pmatrix} \rho_3 & \rho_3 & \rho_3 \\ \rho_3 & \rho_3 & \rho_3 \\ \rho_3 & \rho_3 & \rho_3 \end{pmatrix} \\ 0 & \begin{pmatrix} \rho_1 & \rho_2 & \rho_2 \\ \rho_2 & \rho_1 & \rho_2 \\ \rho_2 & \rho_2 & \rho_1 \end{pmatrix} & 0 \\ \begin{pmatrix} \rho_3 & \rho_3 & \rho_3 \\ \rho_3 & \rho_3 & \rho_3 \\ \rho_3 & \rho_3 & \rho_3 \end{pmatrix} & 0 & \begin{pmatrix} \rho_1 & \rho_2 & \rho_2 \\ \rho_2 & \rho_1 & \rho_2 \\ \rho_2 & \rho_2 & \rho_1 \end{pmatrix} \end{pmatrix}$$

In this example, peak 1 and peak 3 are assumed correlated with prior correlation ρ_3

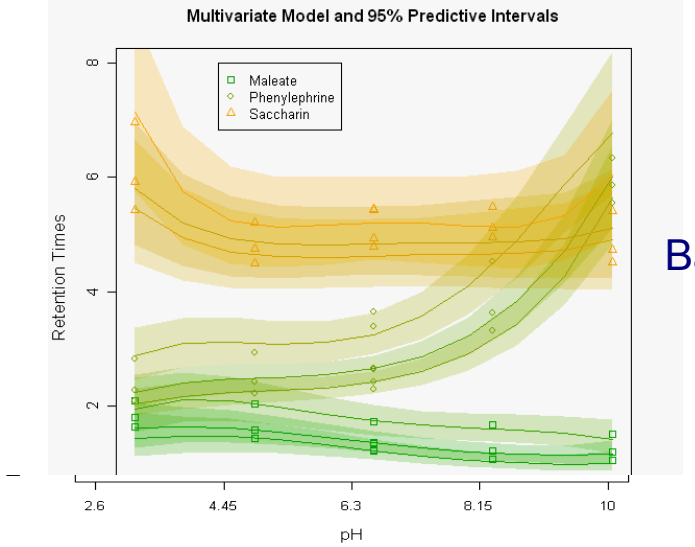
  - When independence can be assumed, it can be interesting to create several independent multivariate models with smaller covariance matrices

Example – Model

- Prior parameters are directly used in the predictive distribution of responses

$$\mathbf{Y}(\mathbf{x}_0)|data \sim T_M\left(\mathbf{M}_{\mathbf{B}\text{post}}\mathbf{z}_0, \left(1+\mathbf{z}_0'(\mathbf{Z}'\mathbf{Z}+\boldsymbol{\Sigma}_0^{-1})^{-1}\mathbf{z}_0\right)\frac{\boldsymbol{\Omega}+\mathbf{A}^*}{\nu+N_0}, \nu+N_0\right)$$

  - Note that if N_0 was set to 0 ($N_0 \geq 0$), The prior Inverse-Wishart distribution would have not been defined ($\nu_0 = N_0 - (M+F)+1$)
  - But the multivariate Student is still defined if $\nu + N_0$ is high enough !



Bayesian predictive intervals (transparent bands) are quantiles of the multivariate Student

Example – Derivation of Quality criteria

- Monte-Carlo simulations allow the (joint) predictive distribution of quality criteria to be propagated from the responses

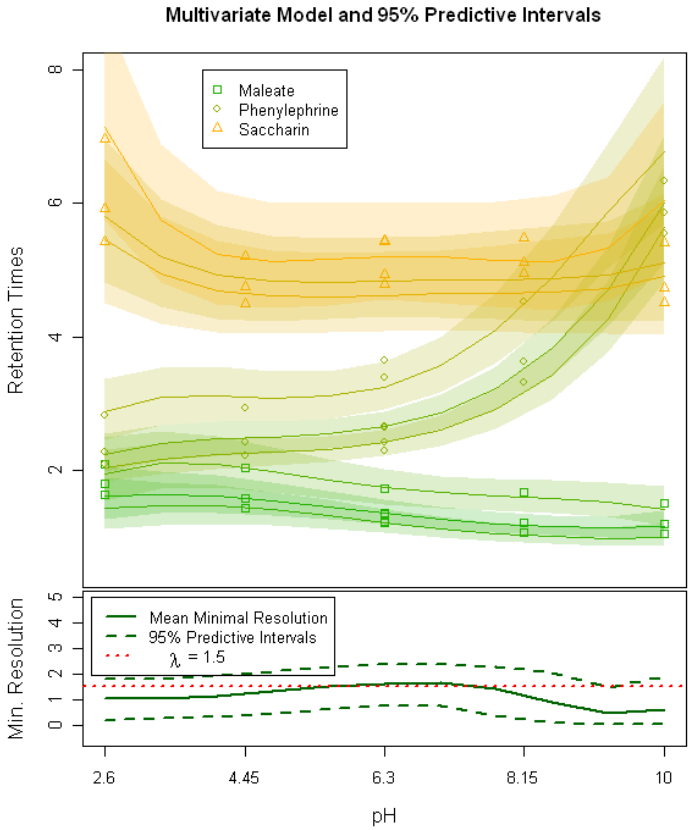
$$\mathbf{Y}(\mathbf{x}_0)|data \sim T_M\left(\ldots \right)$$

$$(B_1,A_1,E_1,\ldots,B_3,A_3,E_3) = g^{-1}(\mathbf{Y}(\mathbf{x}_0)\,|data)$$

Minimal resolution:

$$O_Z = \min_{1\le j\le P-1}\left(\frac{2 * |A_{(j+1)} - A_{(j)}|}{(E_{(j+1)} - B_{(j+1)}) + (E_{(j)} - B_{(j)})}\right)$$

do this for each quality criterion of interest



Bayesian predictive interval (dashed green) is the smallest interval containing $\beta(100)\%$ of the predictive distribution of the criterion

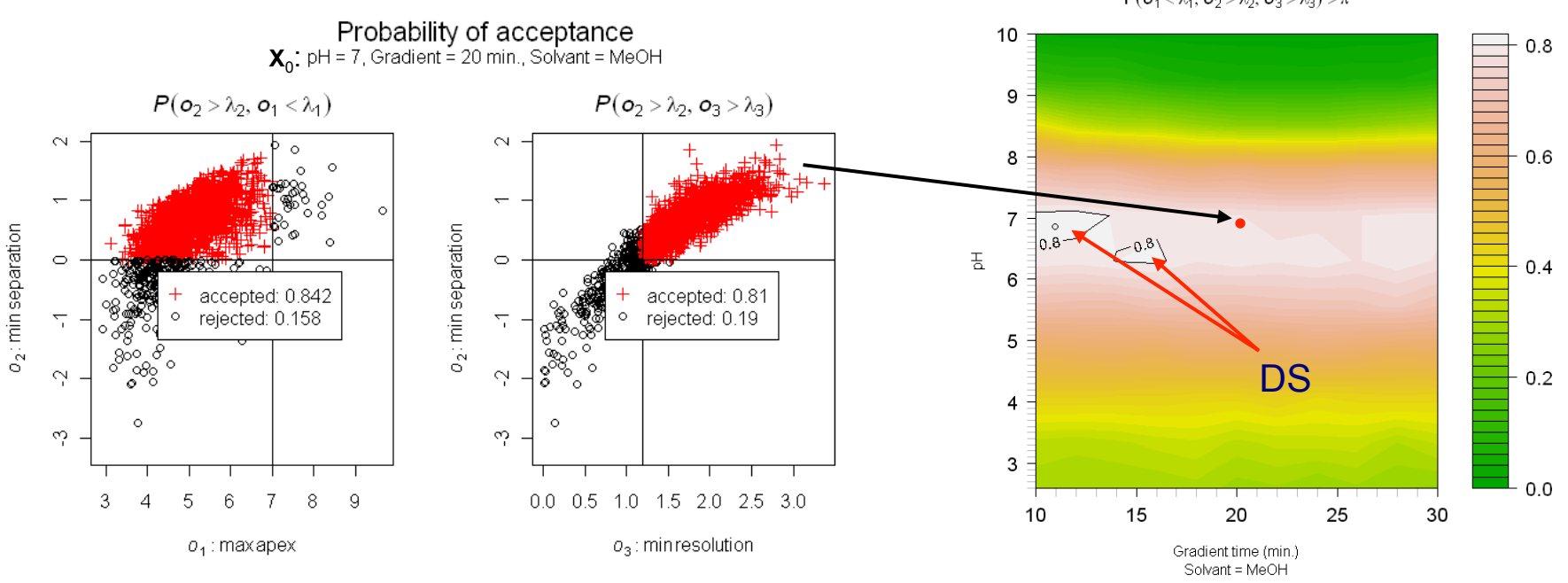
Example – Multi-criteria decision method

- With the joint predictive distribution of criteria, MCDM is made simple !

  - Ex: - separation: $\lambda_1 > 0$ min., resolution: $\lambda_2 > 1.2$, Run : $\lambda_3 < 7$ min. (one-sided)

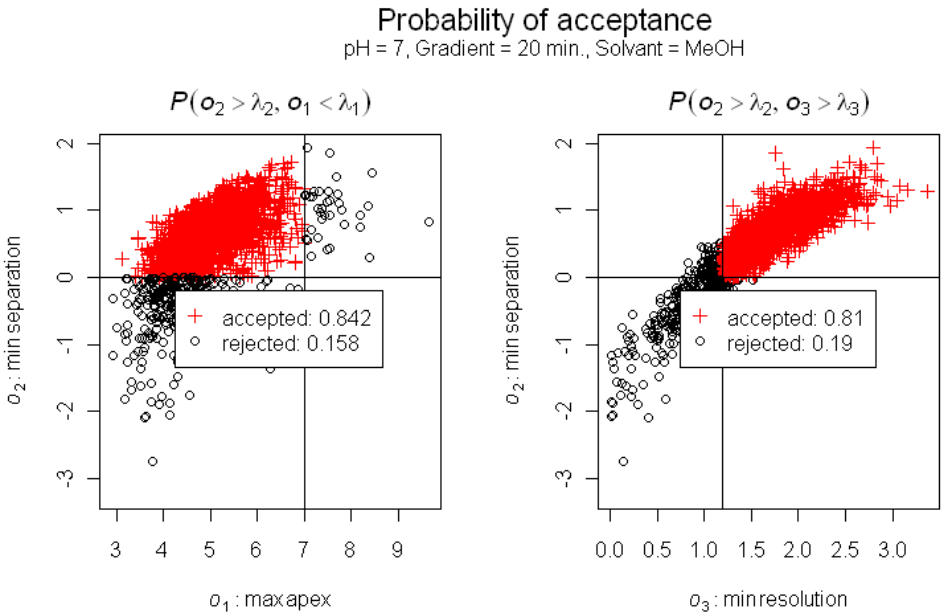
    - quality level : $\pi > 0.8$
    - $E_{\mathbf{B},\boldsymbol{\Sigma}|data}\left[P(\mathbf{O}(\mathbf{x}_0) \in \boldsymbol{\Lambda}) \mid \mathbf{x}_0, \mathbf{B}, \boldsymbol{\Sigma}\right]$?

(Predictive) probability map that the three objectives are achieved



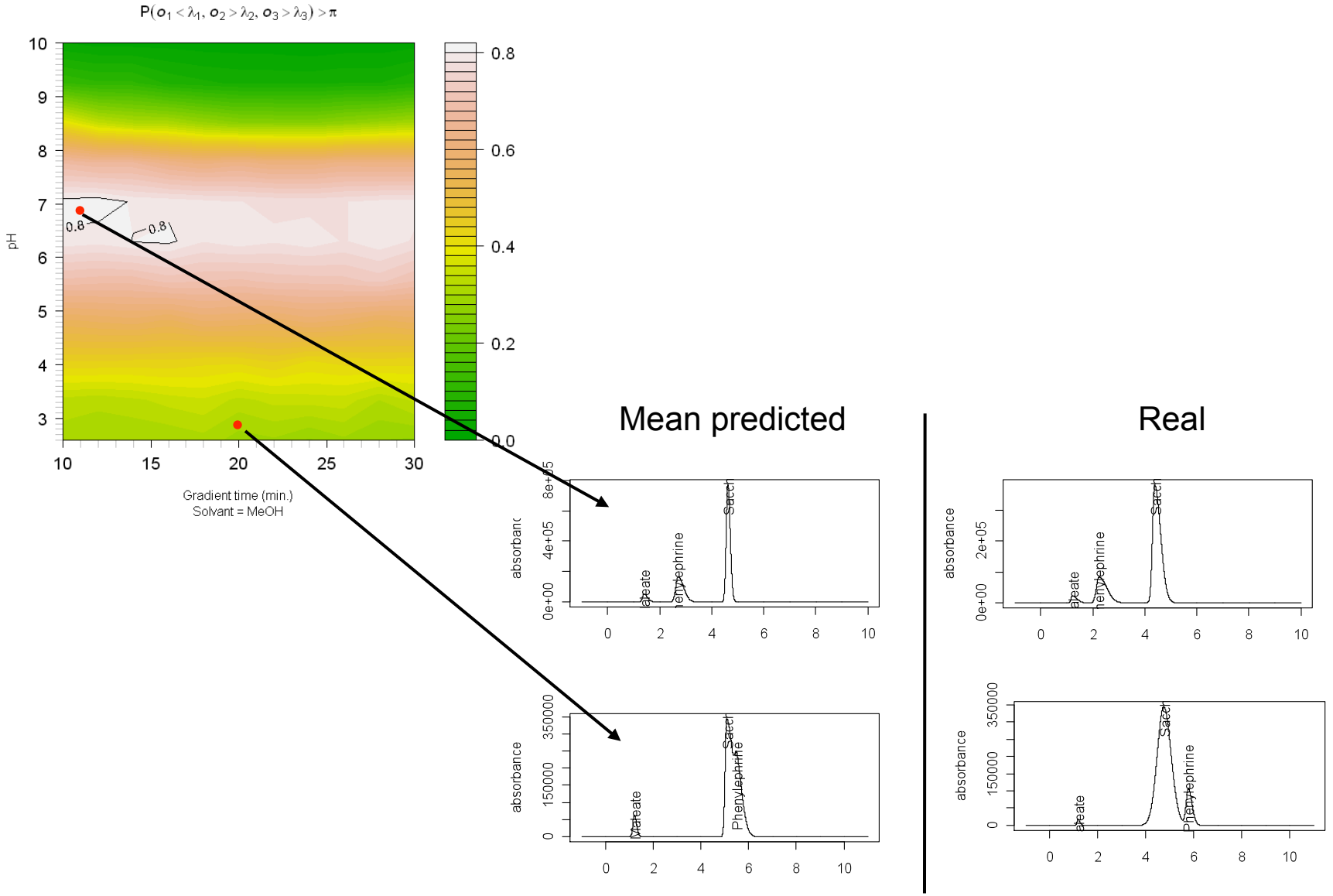
Example – Multi-criteria decision method

- The risk not to achieve the quality criteria is the complementary of the predictive probability to achieve these quality criteria
  - Risk($\mathbf{x}_0$) = 1 - $E_{\mathbf{B},\boldsymbol{\Sigma}|data}\left[P(\mathbf{O}(\mathbf{x}_0) \in \boldsymbol{\Lambda}) \mid \mathbf{x}_0, \mathbf{B}, \boldsymbol{\Sigma}\right]$



Risk = proportion of black points (rejected)

Example – Validation experiments



Conclusions

- We show how it is straightforward to implement Design Space using Bayesian methods
  - With non-informative prior distributions
  - With informative prior distributions
- Bayesian methodology allows finding the predictive distribution of responses
  - → By MCMC simulations or using identified predictive distribution
  - Predict future responses (performance) given past experiments
  - Uncertainty is taken into account
  - …as well as dependencies between responses or quality criteria
- Warning on the possible subjectivity of priors
  - They must be based on past knowledge
  - They should be carefully documented
  - Otherwise, better use non-informative prior distributions

Conclusions

- Bayesian methods provide no help if responses and/or factors are not suited for modelling
  - Classical model checks (residuals, predicted vs. observed, etc.)
  - DIC – Adjusted R^2
  - Known properties of responses (e.g. non linearity)
  - → The simplest model is probably the better
  - → Add a second layer to derive quality criteria if necessary
- If there are reasons to think constraints apply on the responses or criteria. They can be included
  - using truncated distributions (e.g. Geweke, 1991)
  - via rejection sampling, if constraints are complex
- If there are reasons to think (block of) responses can be assumed independent
  - Envisage to model them separately to make several simpler models

- Thanks !
- Any question ?