Overview of Latent Variable Models for Analysis, Optimization & Design

John F. MacGregor

ProSensus, Inc. and McMaster University

Hamilton, ON, Canada

Models

- Learning from data is the key to productivity and quality improvement
- Models in general (mechanistic & theoretical) are simply tools to help us interpret data
- Presentation will focus on empirical Latent Variable models
  - Very powerful for analyzing large volumes of industrial data for:
    - Improved process understanding
    - On-line monitoring
    - Soft sensors
    - Some important control problems (e.g. batch processes)
    - Some important optimization problems
    - Development of new products

Scope

- Multivariate latent variable (LV) methods have been widely used in passive chemometric environments
  - A passive environment is one in which the model is only used to interpret data from a constant environment
    - Calibration
    - Inferential models (soft sensors)
    - Monitoring of processes
- Used much less frequently in an active environment
  - An active environment is one in which the model will be used to actively adjust the process environment
    - Optimization
    - Control
- This talk addresses issues and industrial examples on the use of LV models in both environments.

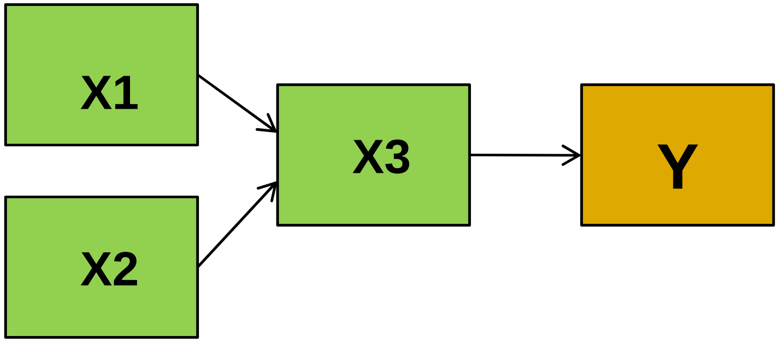
Outline:

- Preliminaries: Some Important Concepts in Latent Variable Modeling for active use
  - Simultaneous modeling of both X & Y spaces
  - Causality of the model
- Passive Applications
  - Analysis of historical data (learning from data)
  - On-line monitoring
- Optimization & Control in Latent Variable spaces
  - Control of final product attributes in batch processes
  - Optimization of processes to achieve desired responses
  - Scale-up and Transfer of products & processes
  - Rapid development of new products

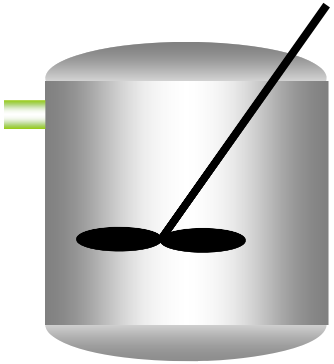
A. Types of Processes and Data Structures

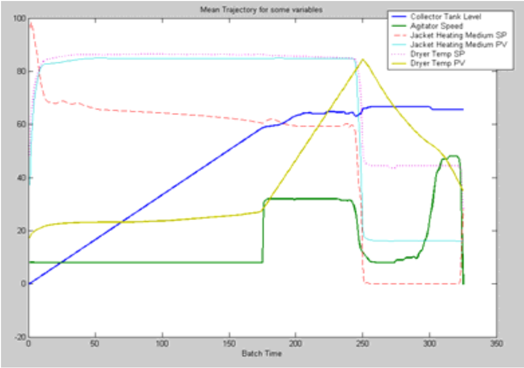
- Continuous Processes

- Data structures

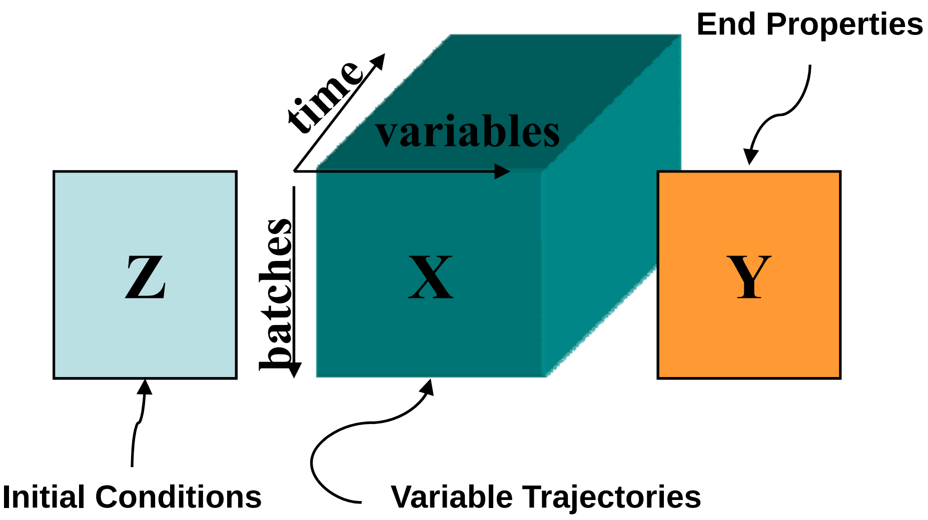


- Batch Processes





- Data structures



Nature of industrial data

- High dimensional data
    - Many variables measured at many times
- Non-causal in nature
    - No cause and effect information among individual variables
- Non-full rank
    - Process really varies in much lower dimensional space
- Missing data
    - 10 – 20 % is common (with some columns/rows missing 90%)
- Low signal to noise ratio
    - Little information in any one variable

- Latent variable models are ideal for these problems

B. Concept of latent variables

Measurements are available on K physical variables: matrix=**X**

K columns

	1	2	3	4	5	6	7	8	9	1
1	Primary ID	Prim In T	Sec In T	Prim Out T	Feed Flow	Chamb P	Diff P Bag/h	System P	Exhaust P	Sec
2	2006-04-05 16:35:00.00	119.049	116.541	41.1646	76.5042	320.199	126.565	66.401	-61.6004	41.
3	2006-04-05 16:35:05.00	119.046	116.532	41.1979	76.4959	325.755	126.636	95.8617	-43.3963	41.
4	2006-04-05 16:35:10.00	119.044	116.523	41.1626	76.4875	321.37	126.708	82.759	-52.5372	41.
5	2006-04-05 16:35:15.00	119.041	116.514	41.1274	76.4792	327.09	126.78	80.6494	-51.5954	41.
6	2006-04-05 16:35:20.00	119.039	116.505	41.101	76.4709	326.797	126.851	94.5307	-43.7692	41.
7	2006-04-05 16:35:25.00	119.036	116.497	41.0367	76.4625	318.052	126.923	85.1925	-50.9631	41.
8	2006-04-05 16:35:30.00	119.034	116.488	41.281	76.4542	323.099	126.995	72.5004	-56.6797	41.

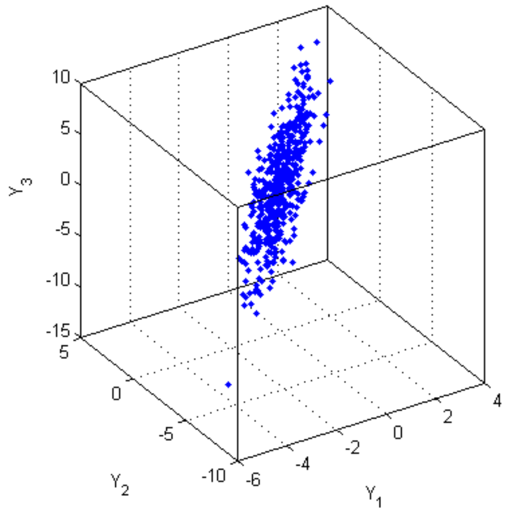
= **X**

But, the process is actually driven by small set of "A" (A « K) *independent* latent variables, called **T.**

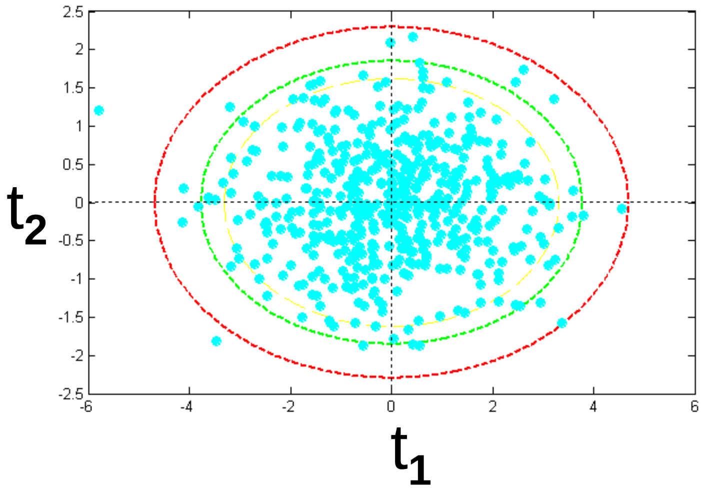
- Raw material variations
- Equipment variations
- Environmental (temp, humidity, etc.) variations

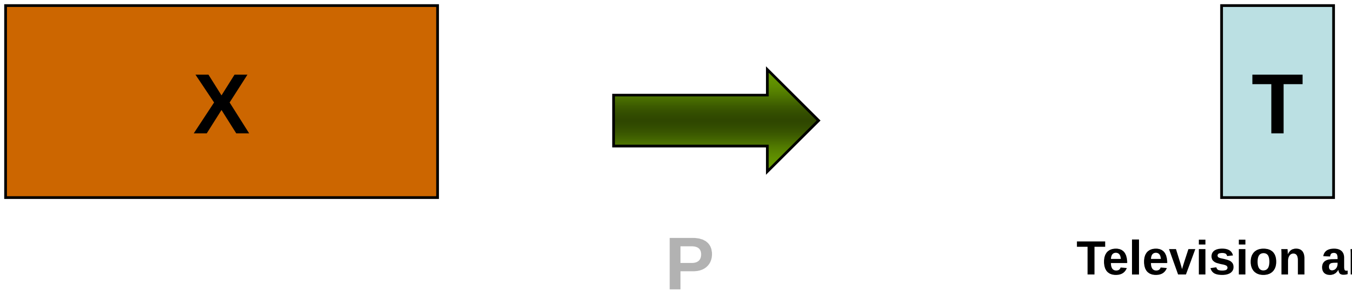
Projection of data onto a low dimensional latent variable space (T)

Measured variables



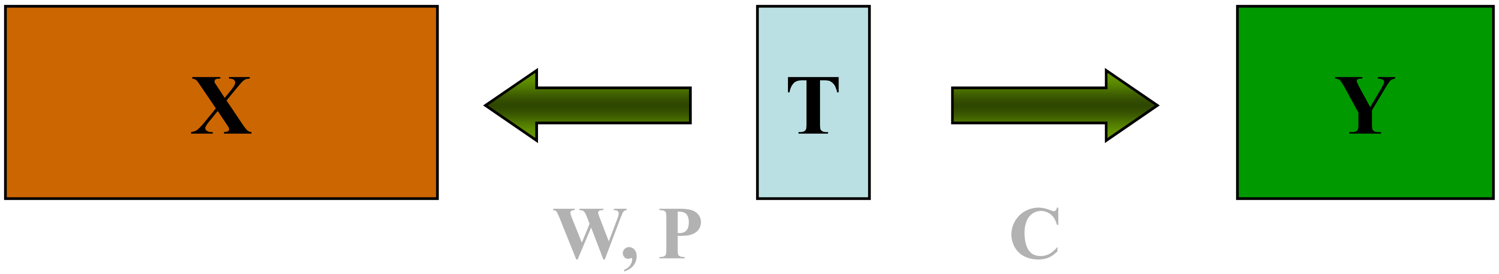
Latent variable space





Television analogy

Latent variable regression models



$$X = TP^T + E$$

$$Y = TC^T + F$$

$$T = XW^*$$

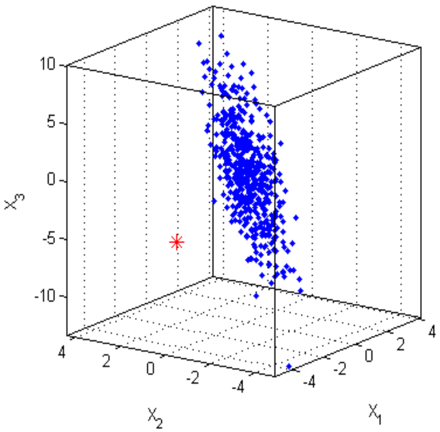
Estimation of W,P,C via PLS

Symmetric in X and Y

- Both X and Y are functions of the latent variables, T
- No hypothesized relationship between X and Y

Important Concepts in Latent Variable Models

- Handle reduced rank nature of the data
  - Work in new low dimensional orthogonal LV space (t1, t2,…)
  - To interpret: Use X model to go back to original variables (contributions)
- Model for X space as well as Y space
  - $X = TP^T + E$; $Y = TC^T + F$
  - Unique among regression methods
  - Essential for uniqueness and for interpretation
  - Essential for checking validity of new data
  - X space model will be the key to all applications in this talk
- Provide causal models in LV space
  - Optimization & control can be done in this space
    - only space where this is justified



Causality in Latent Variable models

- In the passive application of LV models no causality is required
    - Model use only requires that future data follow the same structure
    - No causality is implied or needed among the variables for use of the model
        - Calibration; soft sensors; process monitoring
- For active use such as in optimization and control one needs causal models
    - For empirical models to be causal in certain x-variables – we need to have data with independent variation (DOE's) in those x's.
    - But most process modeling uses "happenstance data" that arise in the natural operation of the process
        - These models do not give causal models for the effect of individual x's on the y's
    - But LV models do provide causal models in the low dimensional LV space
        - Ie. if we move in LV space (t1, t2, …) we can predict the causal effects of these moves on X and Y thru the X and Y space models
        - Will use this fact together with the model of the X-space to perform optimization and control in the LV spaces

C. Industrial illustrations

- Analysis and On-line Monitoring
  - Passive applications

- Control in Latent Variable Spaces:
  - If can monitor on-line, then next step is to take active control action if the batch process is not progressing well.

- Optimization in Latent variable Spaces:
  - Optimization of process operations
  - Scale-up and product transfer between plants
  - Rapid development of new products
  - DOE in LV spaces to improve databases

  - Each example will illustrate the active use of LV models and the importance of working in the LV space and using both X & Y models

C. LV approaches on industrial applications

- Analysis and monitoring of a batch process
- Control of final quality attributes in batch processes
- Optimization of process operation (batch)
- Scale-up and transfer between plants
- Rapid development of new products
- DOE in Latent Variable spaces to enrich dataset.

Analysis of historical data (Process trouble-shooting)

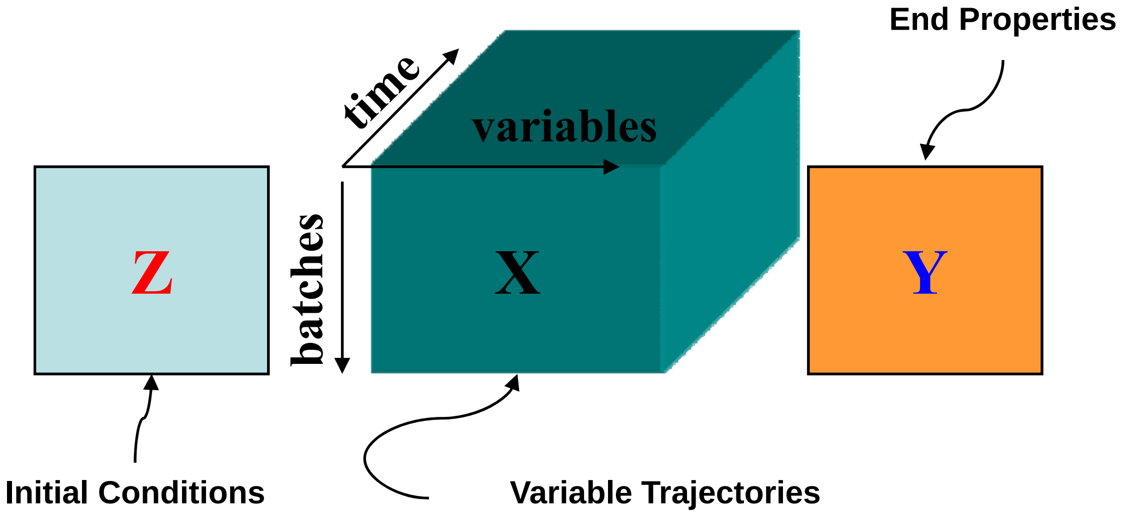
- Where was the process operating poorly?

- Which variables, in which part of the process contribute to this poor behavior?
  - Process understanding

- Industrial example:
  - Herbicide production in a batch manufacturing process

Analysis of an Industrial Agricultural Chemical Process

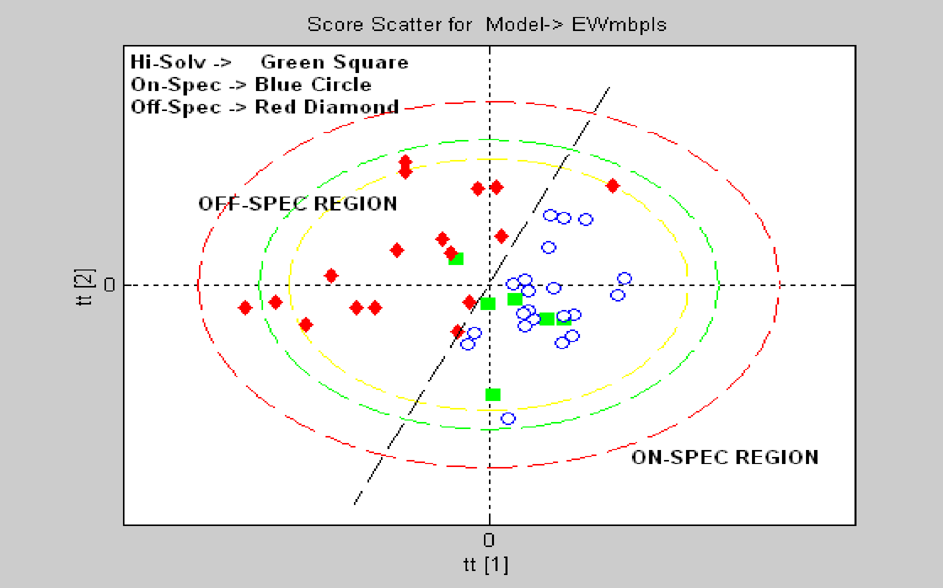
- For each batch (of 72 batches)
  - Raw material properties in **Z** matrix
  - Time varying trajectories of process variables in **X** array
  - Collect product quality data in **Y** matrix



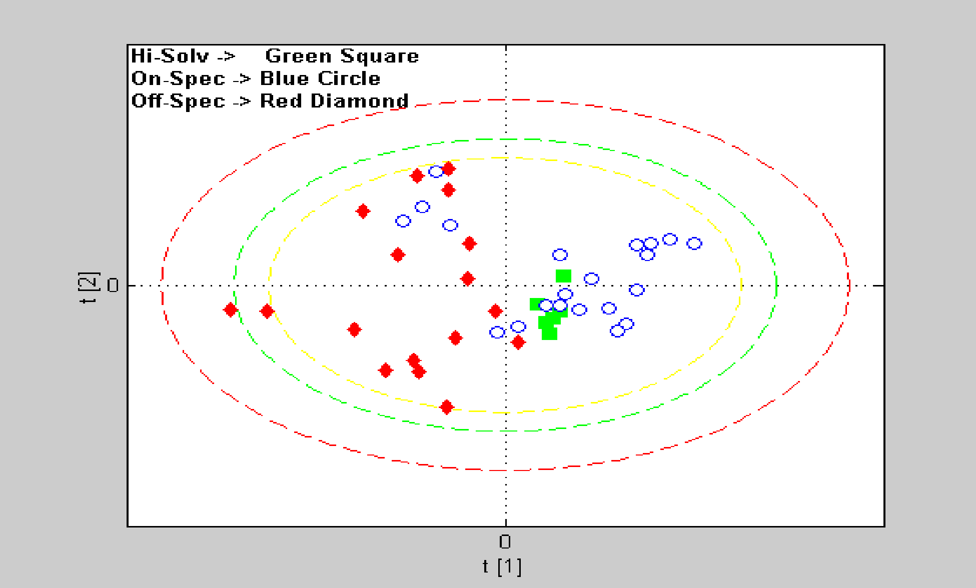
- Database: ~400,000 data points
- PLS latent variable model required only 2 latent variables

ProSensus

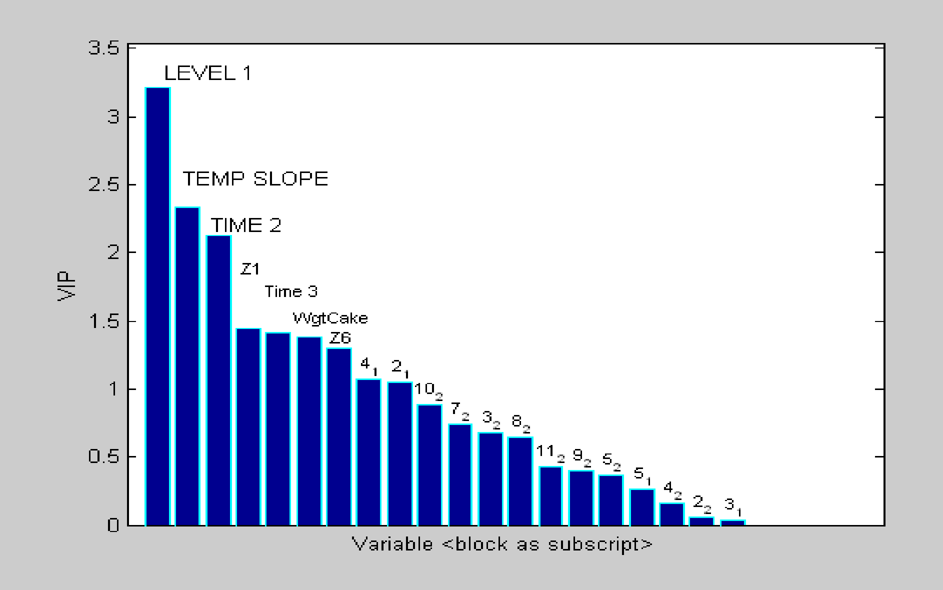
Score plot for Z



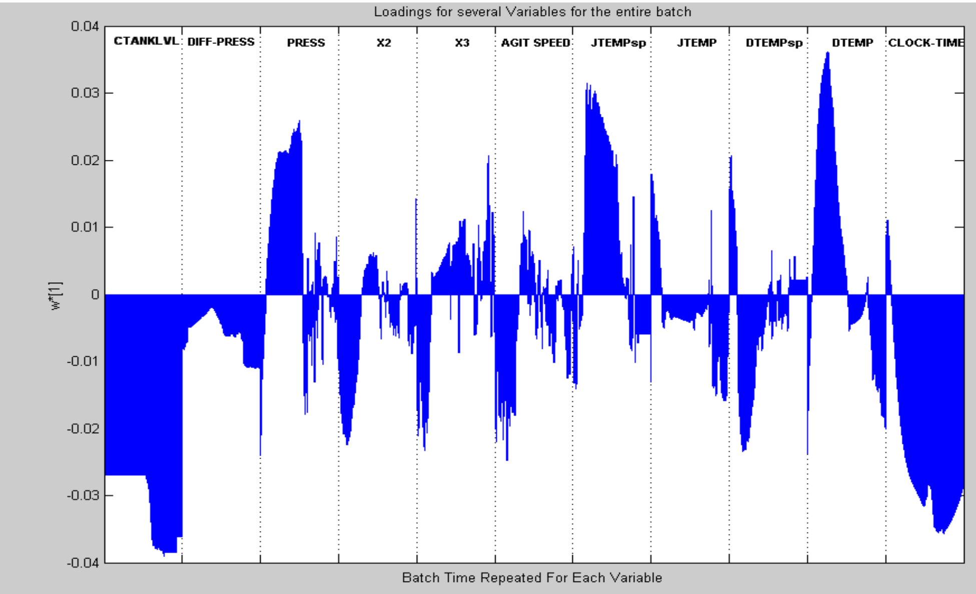
Score lot for X



VIP's for Z model

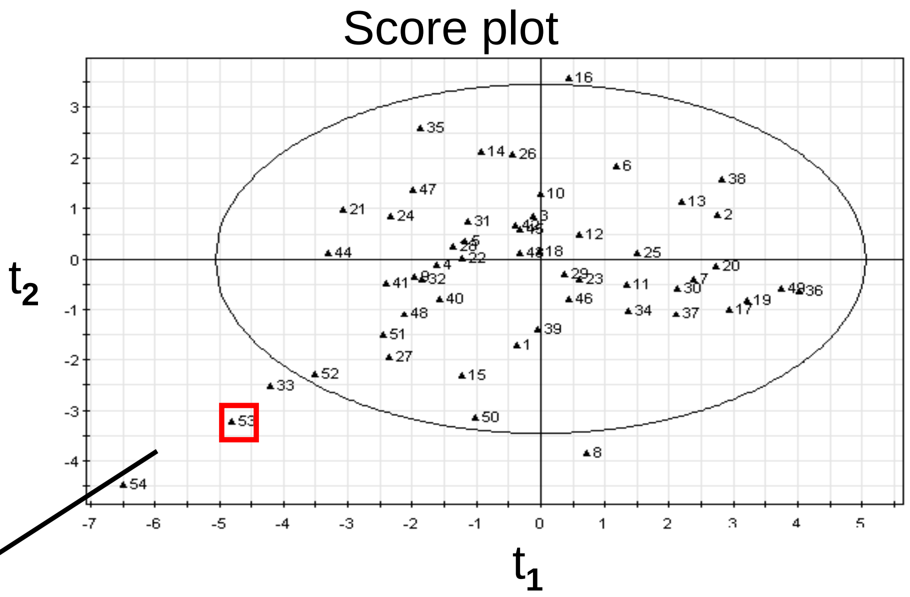


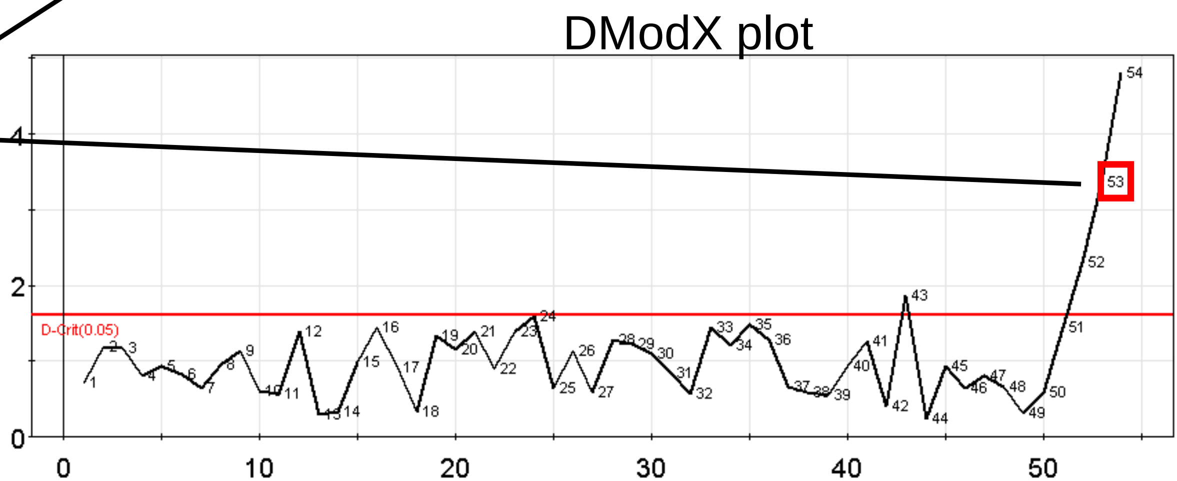
Loading vector w^*_1 for X model



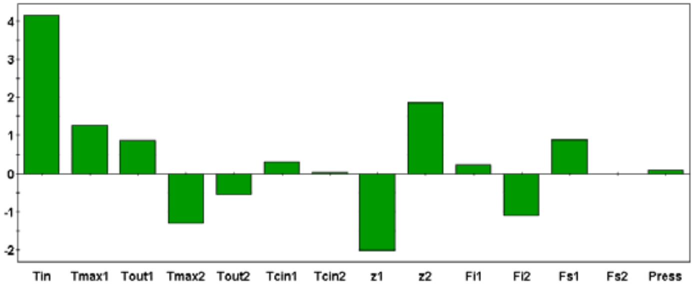
Multivariate process monitoring (MSPC)

- Data: Historical data on process when it has only common cause variation.
- PCA/PLS model
- Monitor new data in the score space of model.
- Are the new data consistent with common cause variation?
- If not, then why not?





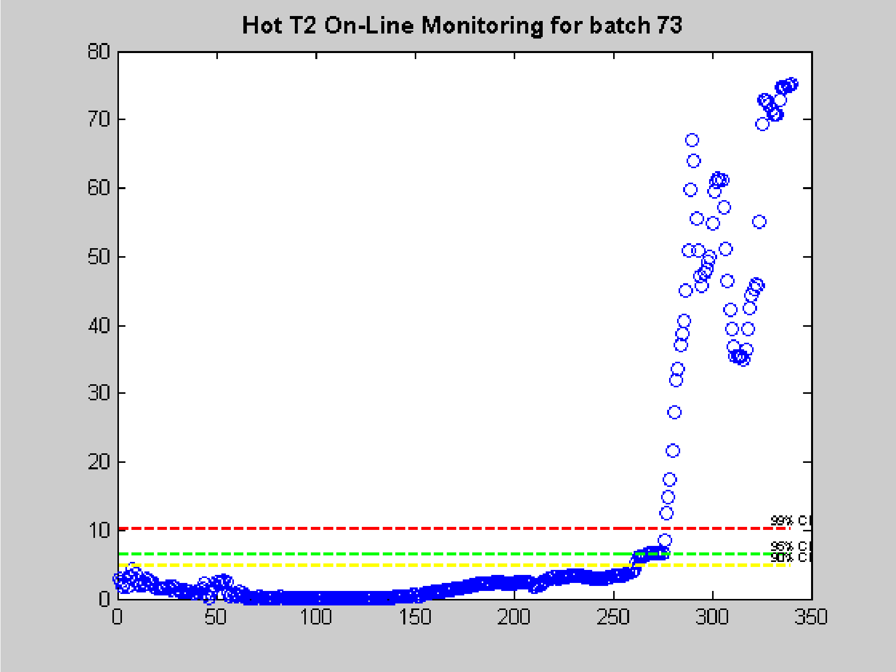
Contribution plots to diagnose



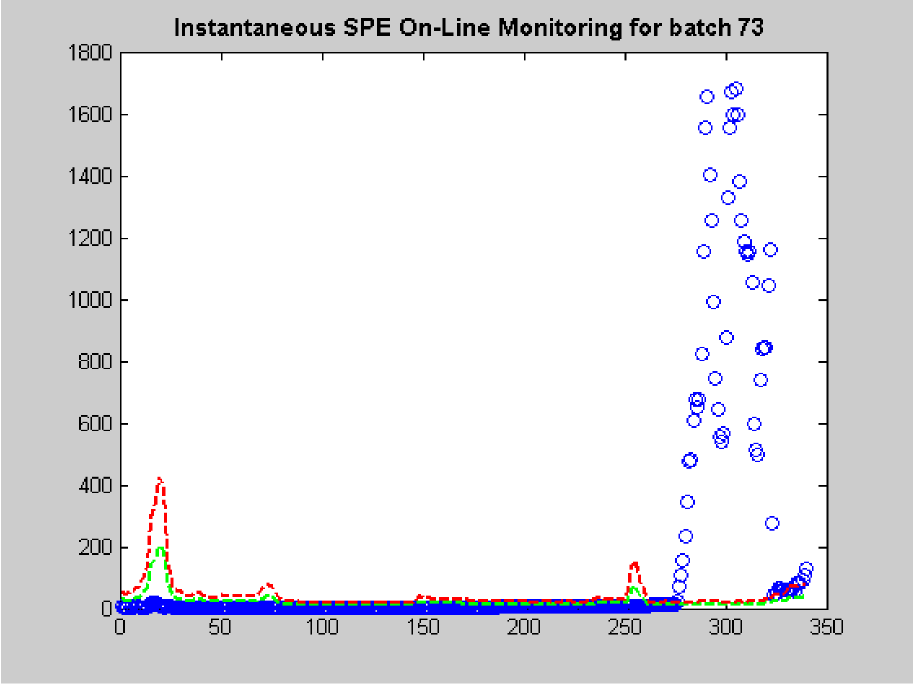
Process monitoring: Agricultural chemical process

Monitoring of new batch number 73

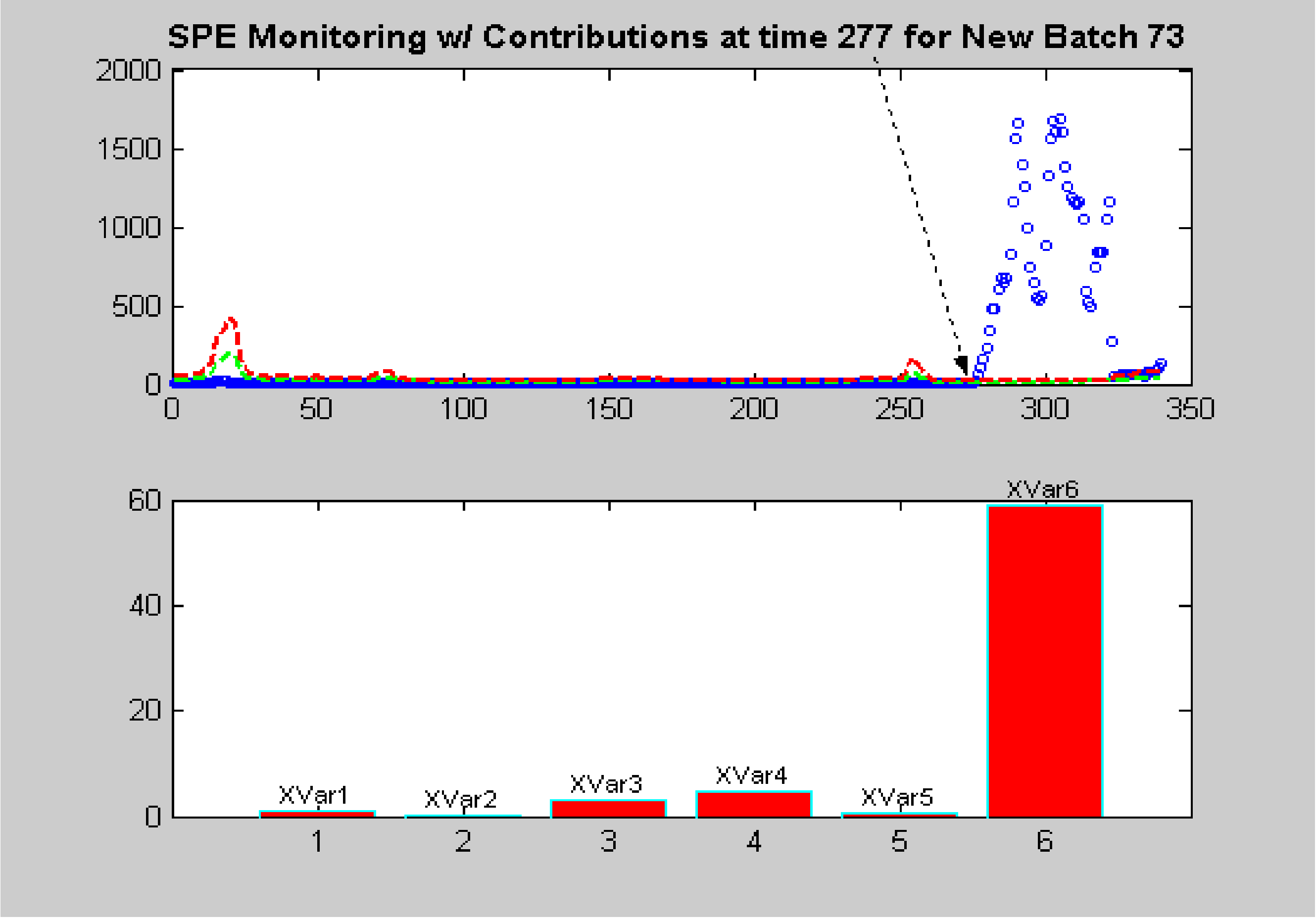
T^2 plot



DMODX plot



Contribution plots to diagnose the problem



Problem: Variable x_6 diverged above its nominal trajectory at time 277

C. LV approaches on industrial applications

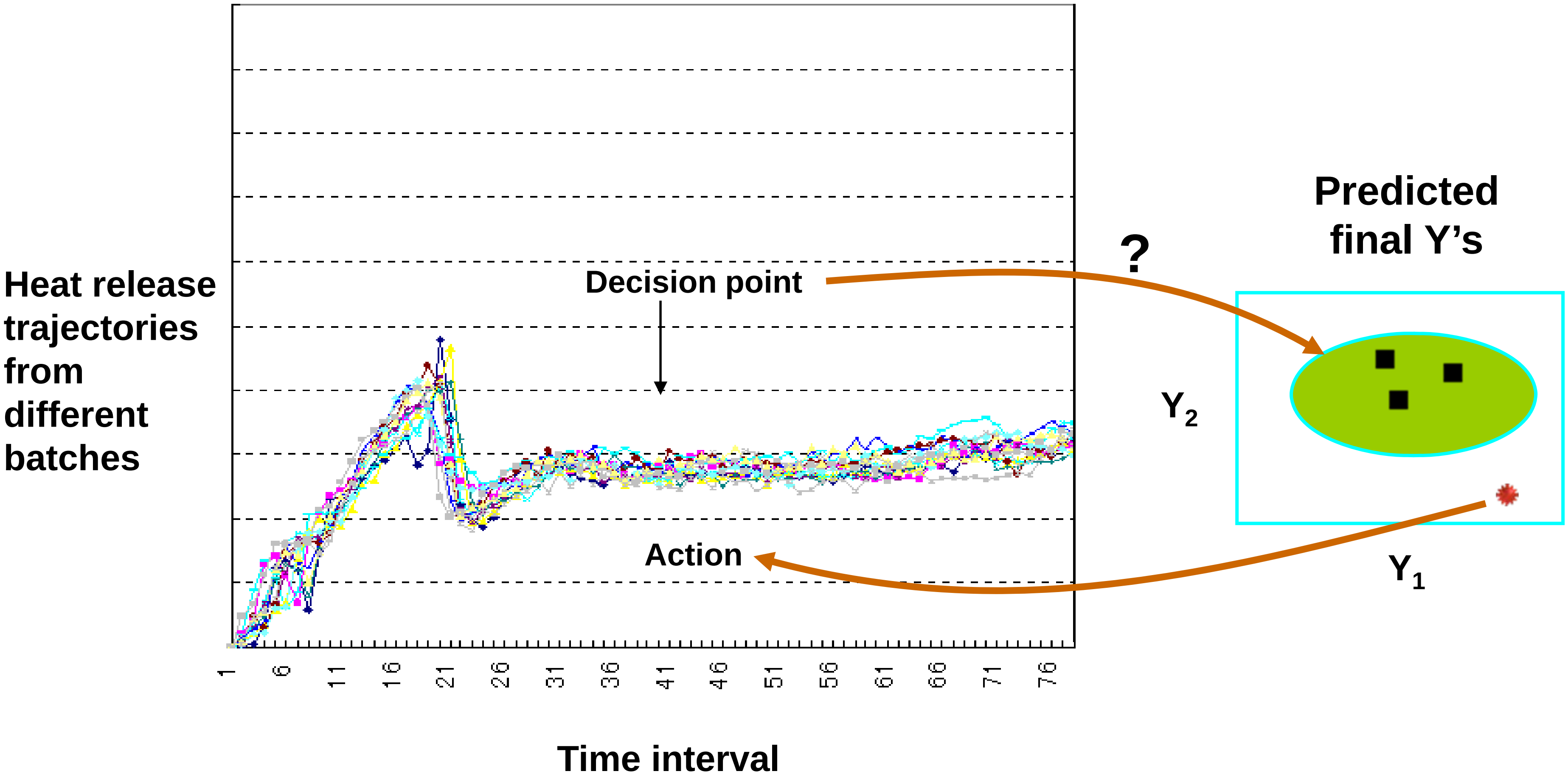
- Analysis and monitoring of a batch process
- Control of final quality attributes in batch processes
- Optimization of process operation (batch)
- Scale-up and transfer between plants
- Rapid development of new products
- DOE in Latent Variable spaces to enrich dataset.

Control of batch product quality

- Objective is to control final product quality
  - e.g. control of final particle size distribution (PSD)
- Using all data up to some decision time, predict final quality with latent variable model
  - All prediction done in low dimensional latent variable space (y's then calculated from t's)
- If predicted quality is outside a desired window, then make a mid-course correction to the batch
  - Control at only one or two points is sufficient
  - Analogy to NASA mid-course rocket trajectory adjustment in moon missions
- Data requirement: Historical batches + few with DOE on corrective variables

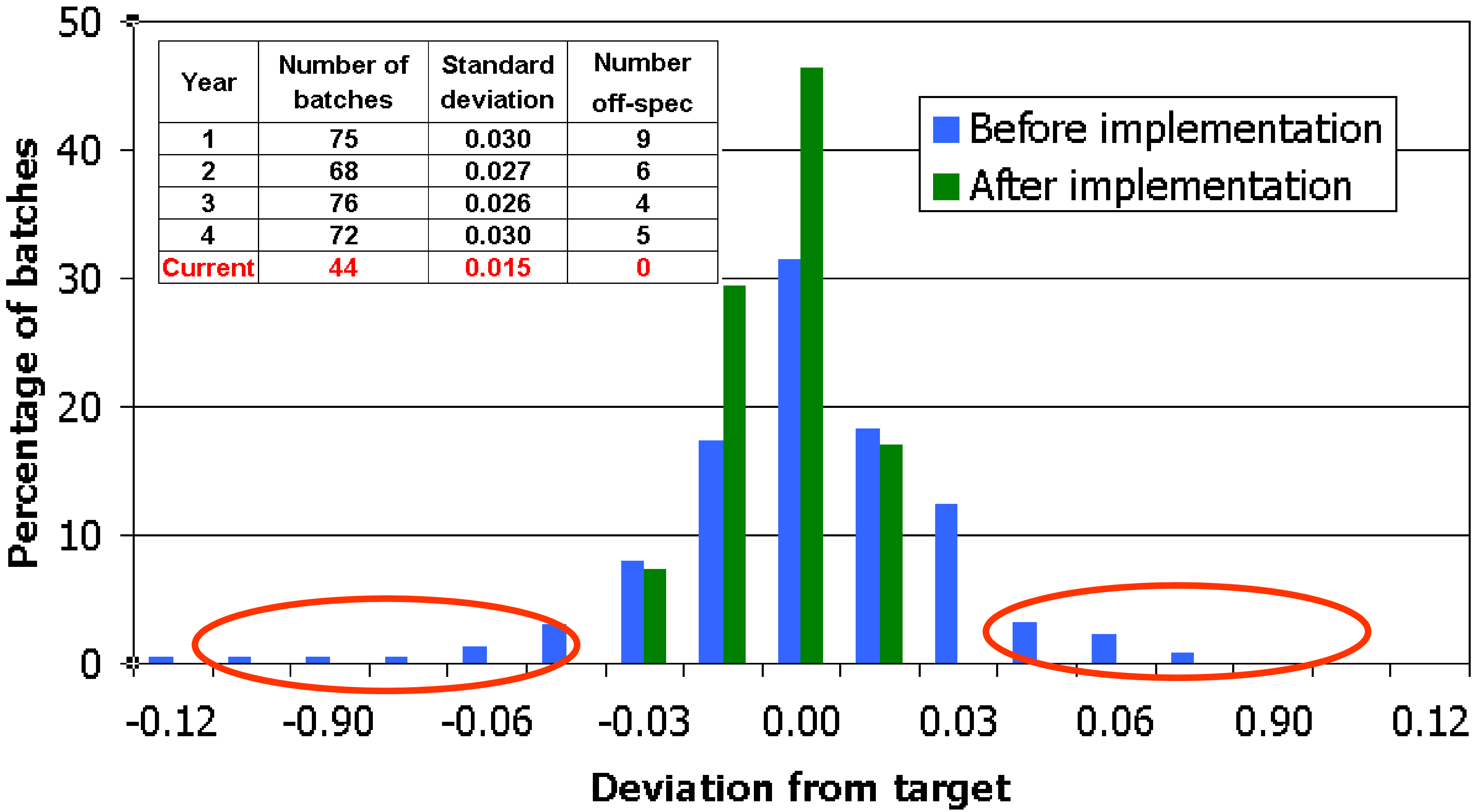
Control of PSD via mid-course correction

- Trajectories of one variable from many batches
- At decision point – predict Y's – if outside target region – take action



Good industrial results (Mitsubishi Chemicals)

Mid-course control : before and after implementation

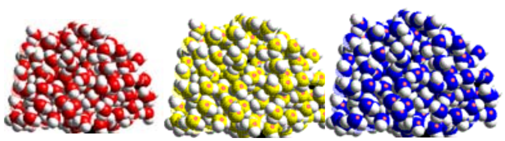


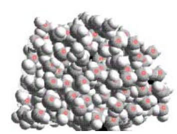
Year	Number of batches	Standard deviation	Number off-spec
1	75	0.030	9
2	68	0.027	6
3	76	0.026	4
4	72	0.030	5
Current	44	0.015	0

C. LV approaches on industrial applications

- Analysis and monitoring of a batch process
- Control of final quality attributes in batch processes
- Optimization of process operation (batch)
- Scale-up and transfer between plants
- Rapid development of new products
- DOE in Latent Variable spaces to enrich dataset.

Optimizing operating policies for new products





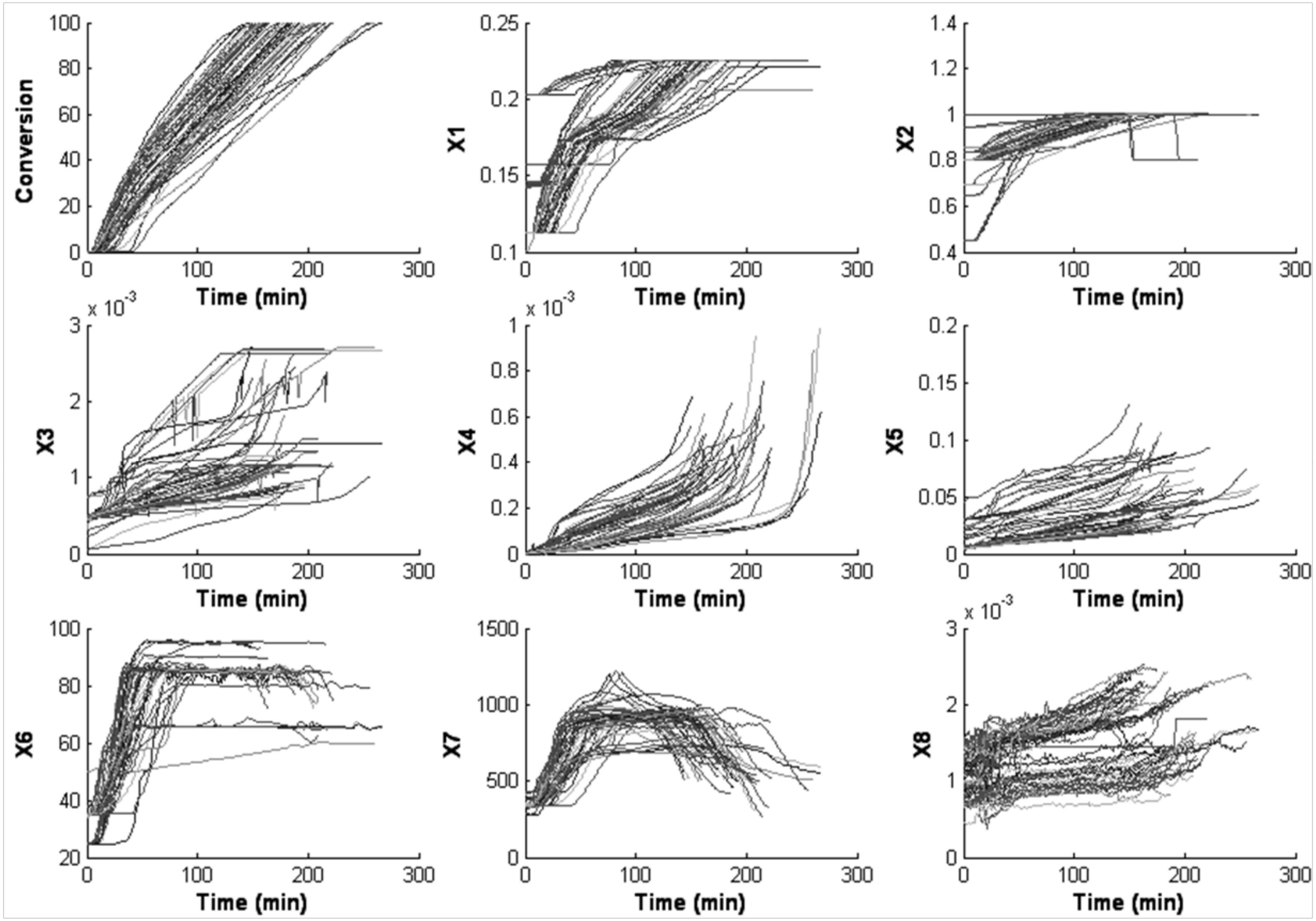
X

- Temperatures
- Pressures
- Concentrations
- Recipes
- Flows
- Trajectories

Y

- Density
- Tensile strength
- Mw, Mn
- Transparency
- Biological activity
- Toxicity
- Hydrophobicity

Batch polymerization process trajectory data (X)



Batch emulsion polymerization (Air Products & Chemicals)

- 13 variables in Y

  Desire a new product with the following final quality attributes (Y's):

  Maintain in normal ranges: $Y_1 \quad Y_2 \quad Y_3 \quad Y_4 \quad Y_5 \quad Y_6 \quad Y_8$

  Constraints:

  $Y_7 = Y_{7des}$
  $Y_9 = Y_{9des}$
  $Y_{10} < Y_{10const}$
  $Y_{11} < Y_{11const}$
  $Y_{12} < Y_{12const}$
  $Y_{13} < Y_{13const}$

  … and with the minimal possible batch time (*)

- Solution
  - Build batch PLS latent variable model on existing data (Z, X, Y)
  - Perform an optimization in LV space to find optimal LV's
  - Use LV model of X-space to find the corresponding recipes and process trajectories

Process Optimization

- Design via PLS model inversion (no constraints)

PLS Model:

$$\hat{\mathbf{Y}} = \mathbf{T}\mathbf{Q}^{\mathbf{T}}$$

$$\hat{\mathbf{y}}_{\mathbf{des}} = \mathbf{Q}\,\boldsymbol{\tau}_{\mathbf{new}}$$

Step 1

$$\tau_{new} = inv\left(Q^T Q\right) Q^T y_{des}$$

$$\hat{\mathbf{X}} = \mathbf{T}\mathbf{P}^{\mathbf{T}}$$

$$\hat{\mathbf{x}}_{\mathbf{new}} = \mathbf{P}\,\boldsymbol{\tau}_{\mathbf{new}}$$

Step 2

- If dim(Y) < dim(X) then is a null space
  - A whole line or plane of equivalent solutions yielding the same y_{des}

Solution with constraints: Formulate inversion as an optimization

- **Step 1**: Solve for $\hat{\tau}_{new}$ with constraints on T² and on y's

$$\min_{\hat{\boldsymbol{\tau}}_{xnew}} \left\{ (\mathbf{y}_{\text{des}} - \mathbf{Q}\,\hat{\boldsymbol{\tau}}_{\text{xnew}})^{\mathrm{T}} \mathbf{G}_1 (\mathbf{y}_{\text{des}} - \mathbf{Q}\,\hat{\boldsymbol{\tau}}_{\text{xnew}}) + \rho \left(\sum_{a=1}^{A} \frac{\hat{\tau}^2_{xnew,a}}{s_a^2} \right) \right\}$$

$$s.t$$

$$\mathbf{B}\,\mathbf{Q}\,\hat{\boldsymbol{\tau}}_{\mathbf{xnew}} < \mathbf{b}$$

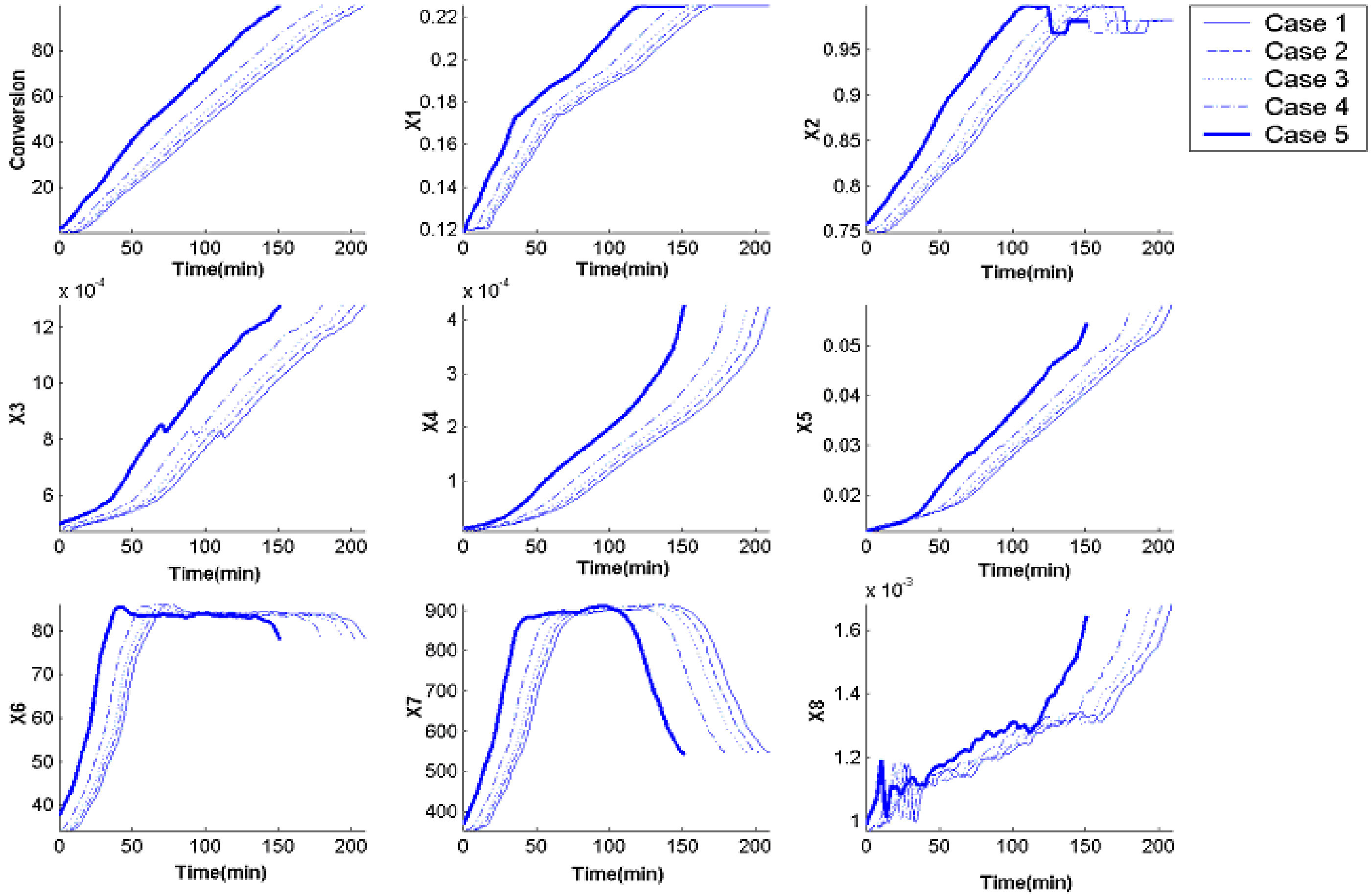
- **Step 2**: Solve for x_{new} that yields $\hat{\tau}_{new}$ subject to certain constraints on SPE and x's.

$$\min_{\mathbf{x}_{\text{new}}} \left\{ \left(W^* \mathbf{x}_{\text{new}} - \hat{\boldsymbol{\tau}}_{\text{new}} \right)^{\mathrm{T}} \mathbf{G}_2 \left(W^* \mathbf{x}_{\text{new}} - \hat{\boldsymbol{\tau}}_{\text{new}} \right) + \left(\mathbf{x}_{\text{new}} - PW^* \mathbf{x}_{\text{new}} \right)^{\mathrm{T}} \boldsymbol{\Lambda} \left(\mathbf{x}_{\text{new}} - PW^* \mathbf{x}_{\text{new}} \right) + \boldsymbol{\eta} \mathbf{x}_{\text{new}} \right\}$$

Different solutions: change the penalty (η) on time usage

All solutions satisfy the requirements on y_{des}

Case 1 to 5: weight on time-usage is gradually increased

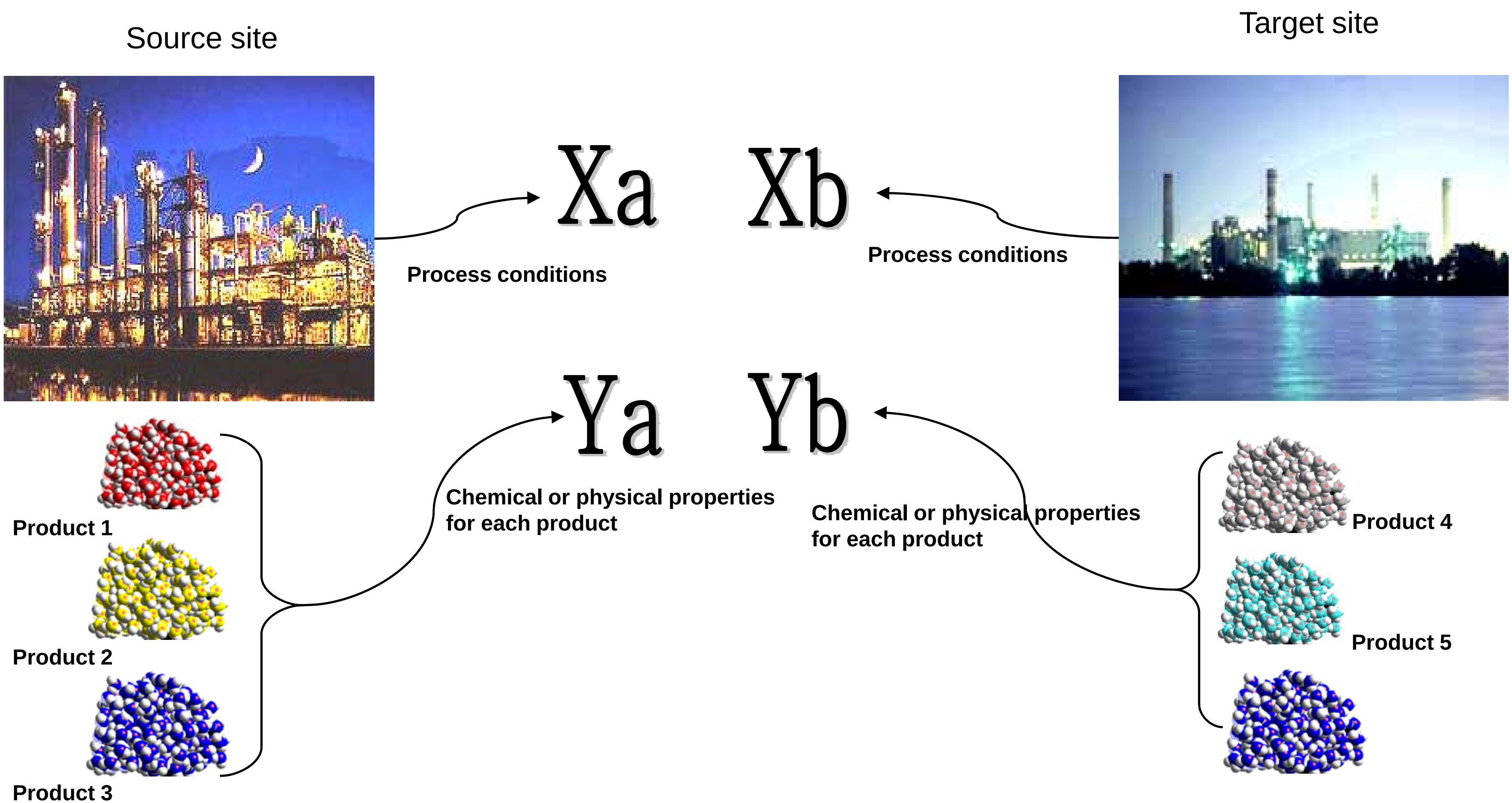


Garcia-Munoz, S., J.F. MacGregor, D. Neogi, B.E. Latshaw and S. Mehta, "Optimization of batch operating policies. Part II: Incorporating process constraints and industrial applications", Ind. & Eng. Chem. Res., 2008

C. LV approaches on industrial applications

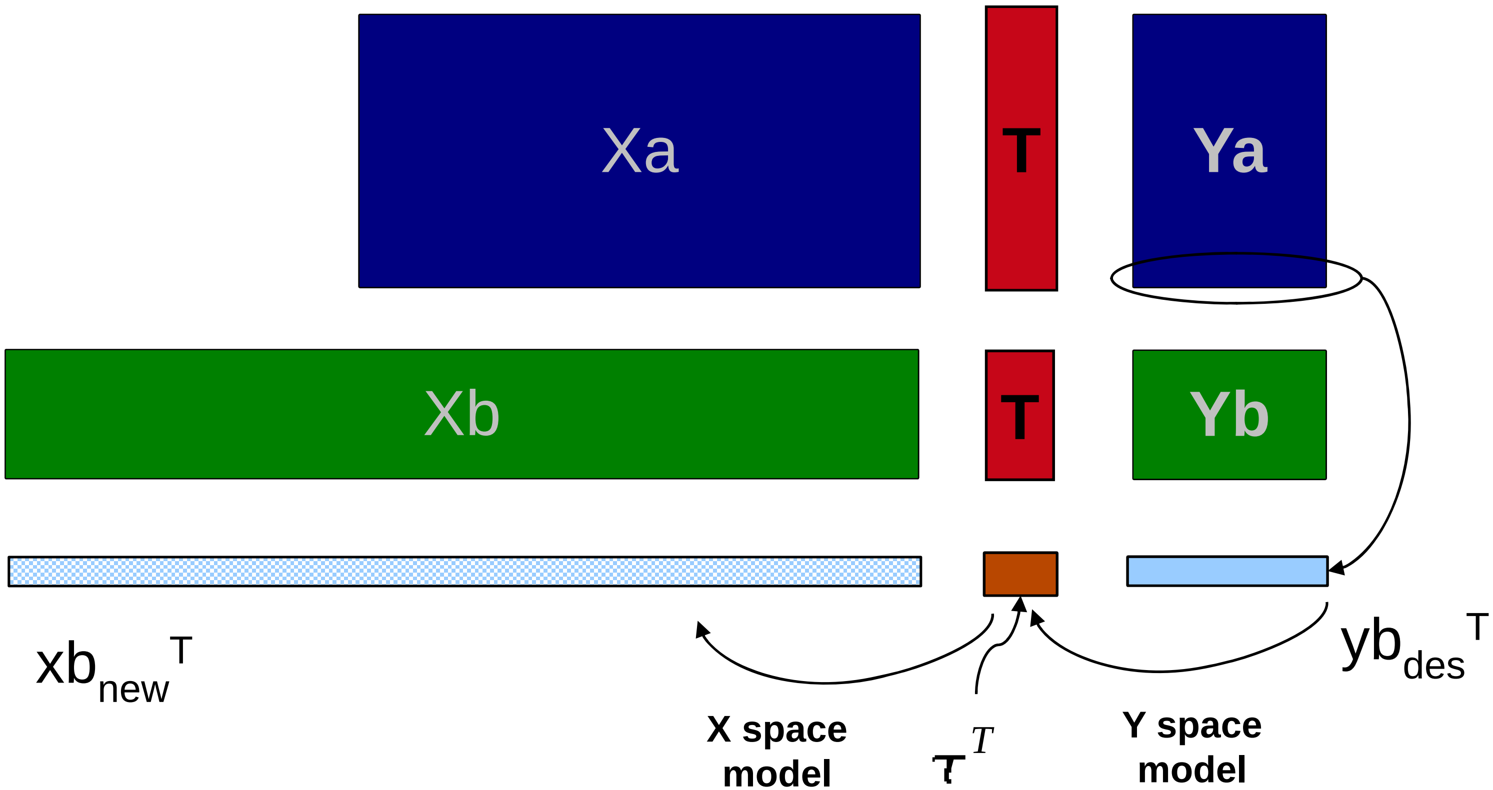
- Analysis and monitoring of a batch process
- Control of final quality attributes in batch processes
- Optimization of process operation (batch)
- Scale-up and transfer between plants
- Rapid development of new products
- DOE in Latent Variable spaces to enrich dataset.

Product transfer between plants and scale-up



Product transfer and scale-up

Historical data from the 2 plants. Build JYPLS model

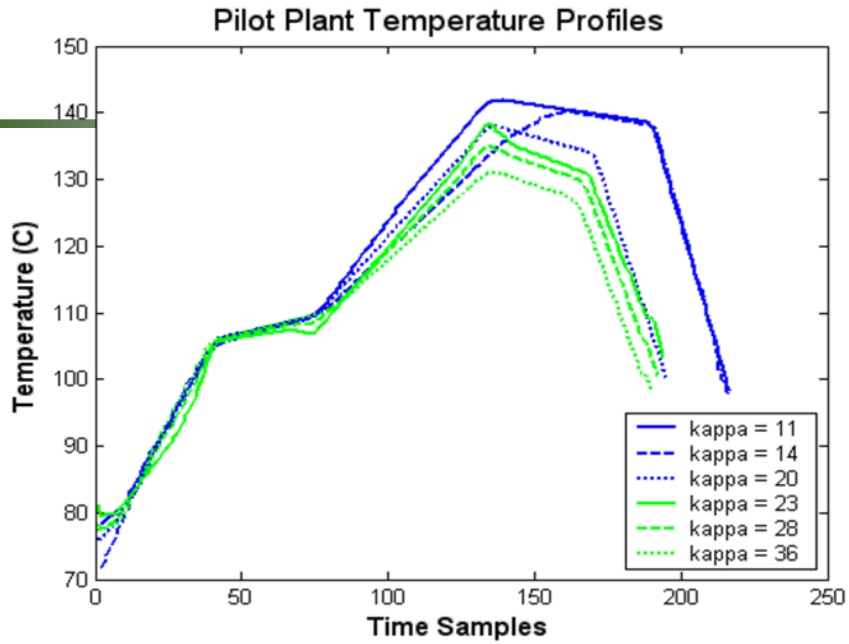


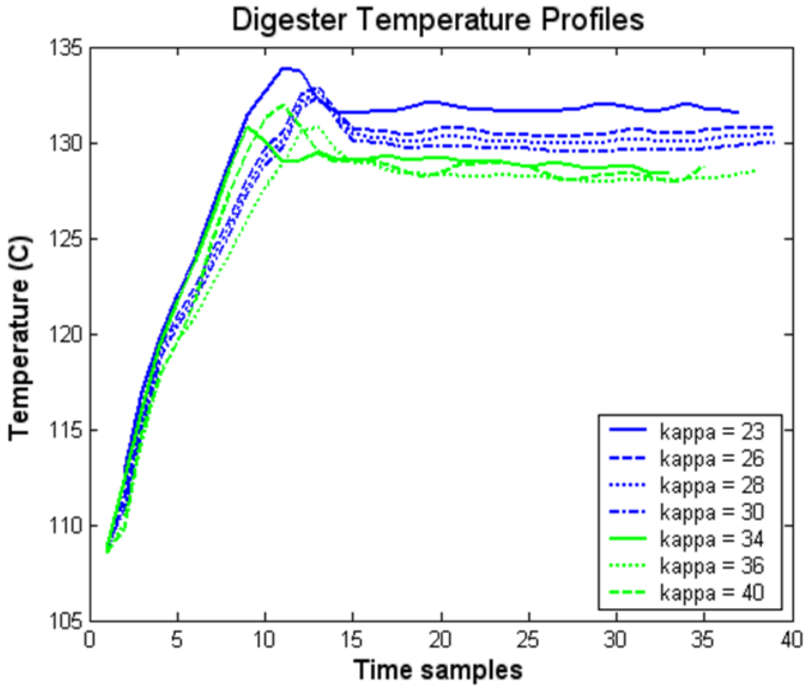
Garcia-Munoz, S., T.Kourti and J.F. MacGregor, “Product Transfer Between Sites using Joint-Y PLS”, Chemometrics & Intell. Lab. Systems, 79, 101-114, 2005.

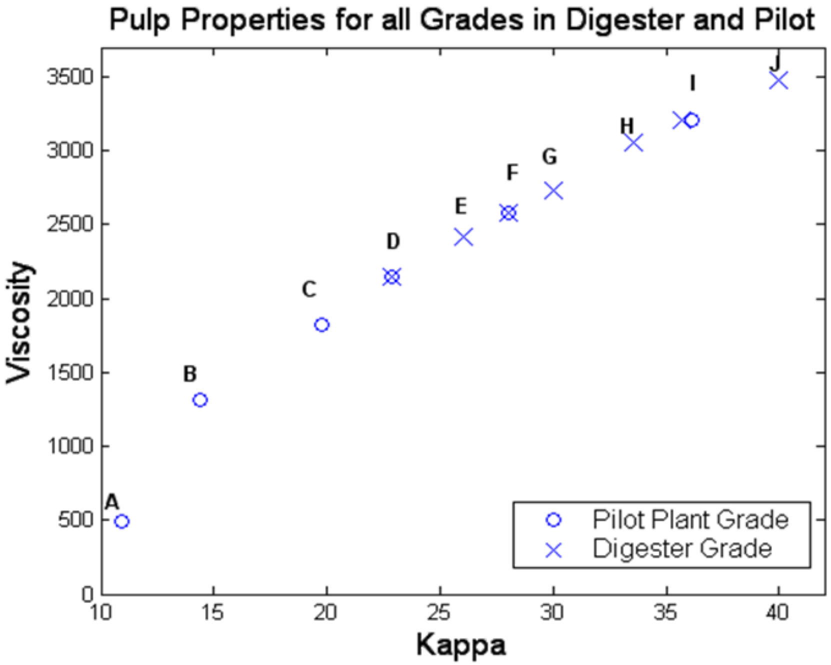
Industrial Scale-up Example

Tembec - Cdn. pulp & paper company:

Pilot plant and full scale digesters



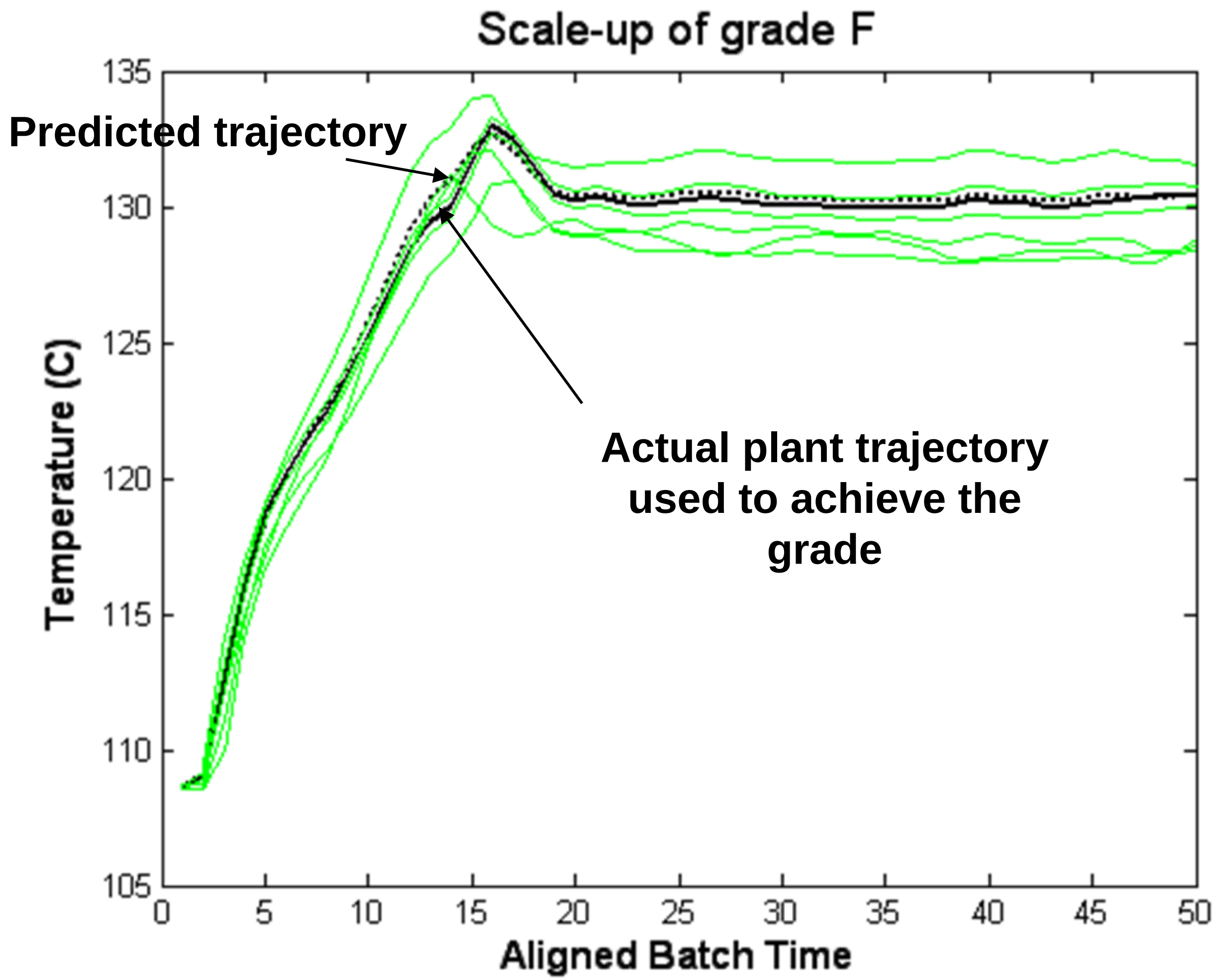




Scale up for grade F – pulp digester

Build models on all pilot plant data and all plant data (ex F)

Design operating profiles to achieve grade F in plant.



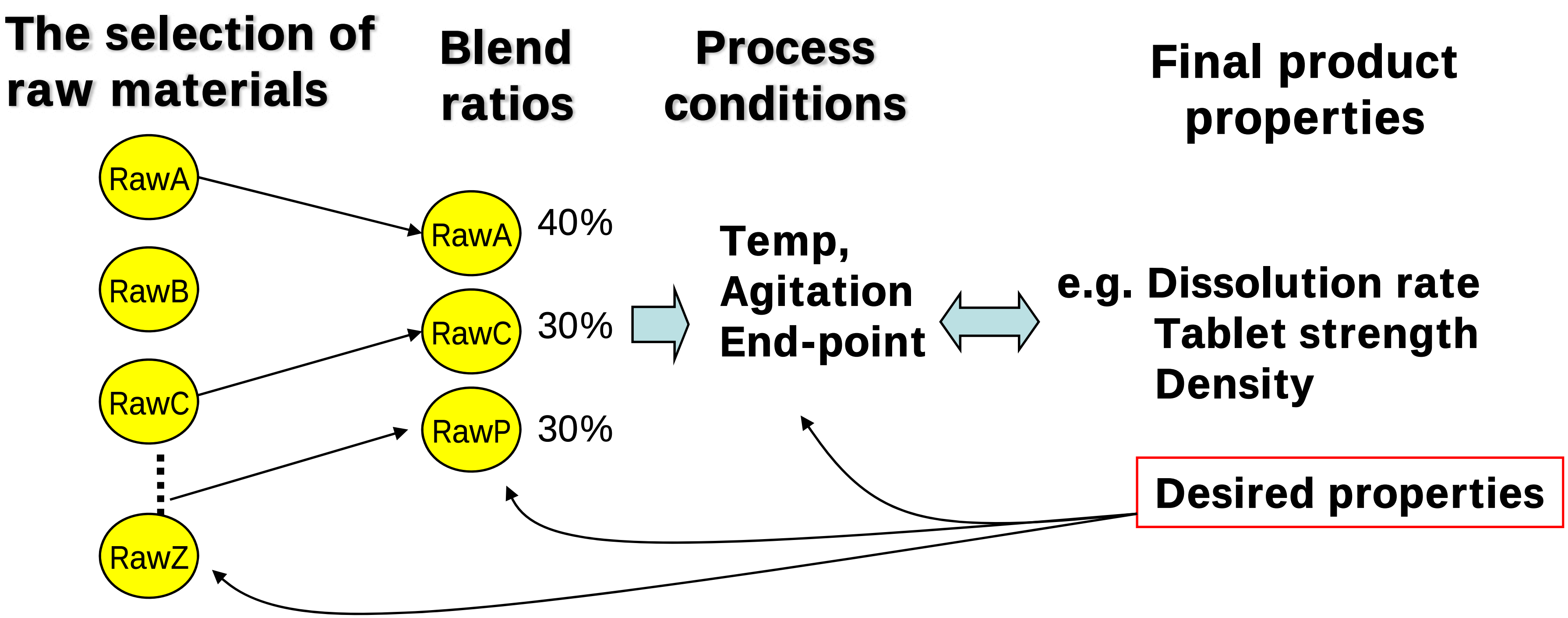
C. LV approaches on industrial applications

- Analysis and monitoring of a batch process
- Control of final quality attributes in batch processes
- Optimization of process operation (batch)
- Scale-up and transfer between plants
- Rapid development of new products
- DOE in Latent Variable spaces to enrich dataset.

Rapid Development of New Products

- Companies accumulate lot of data on their products and processes.
- Can we use that data to rapidly develop new products?
- Three general degrees of freedom for developing new products:
  - Raw material selection
  - Ratios in which to use raw materials (formulation)
  - Process conditions for manufacturing
    - Relative importance of these three depends on the industry and the product
    - Huge synergisms among these

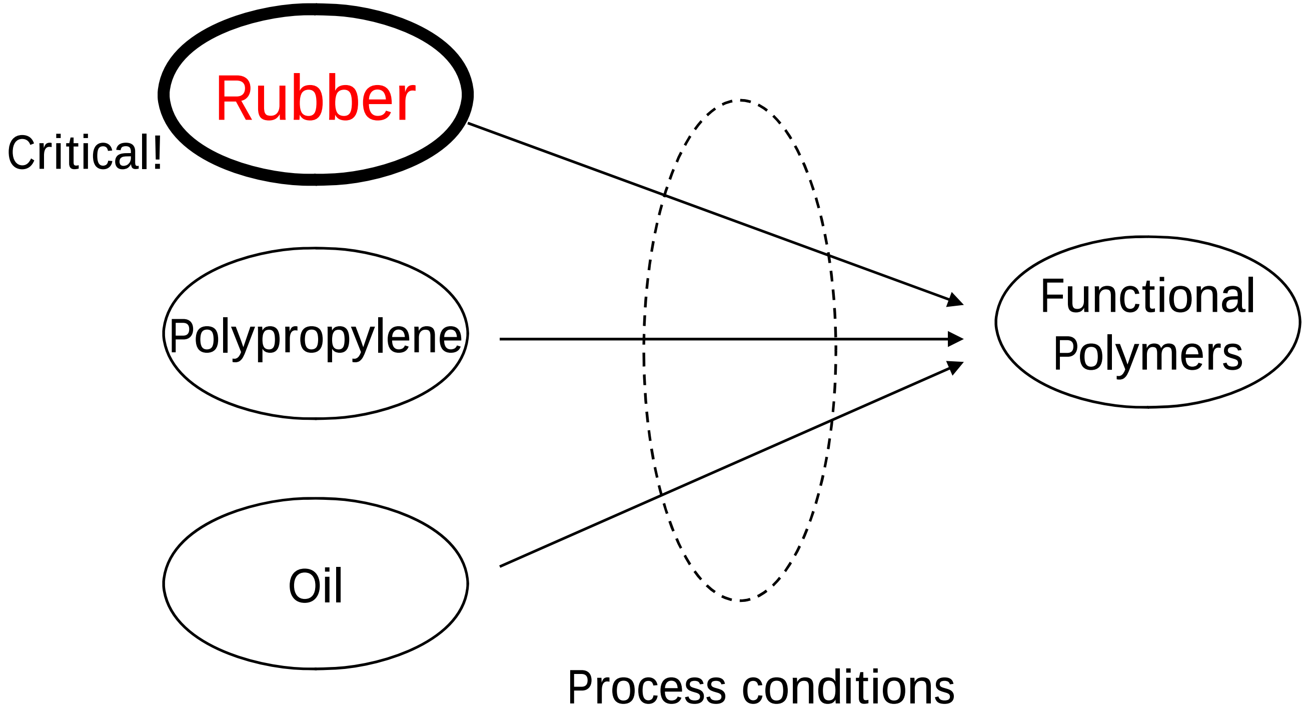
What is the problem ?



Traditional approaches tend to treat each step separately→ inefficient as they miss synergism among these degrees of freedom

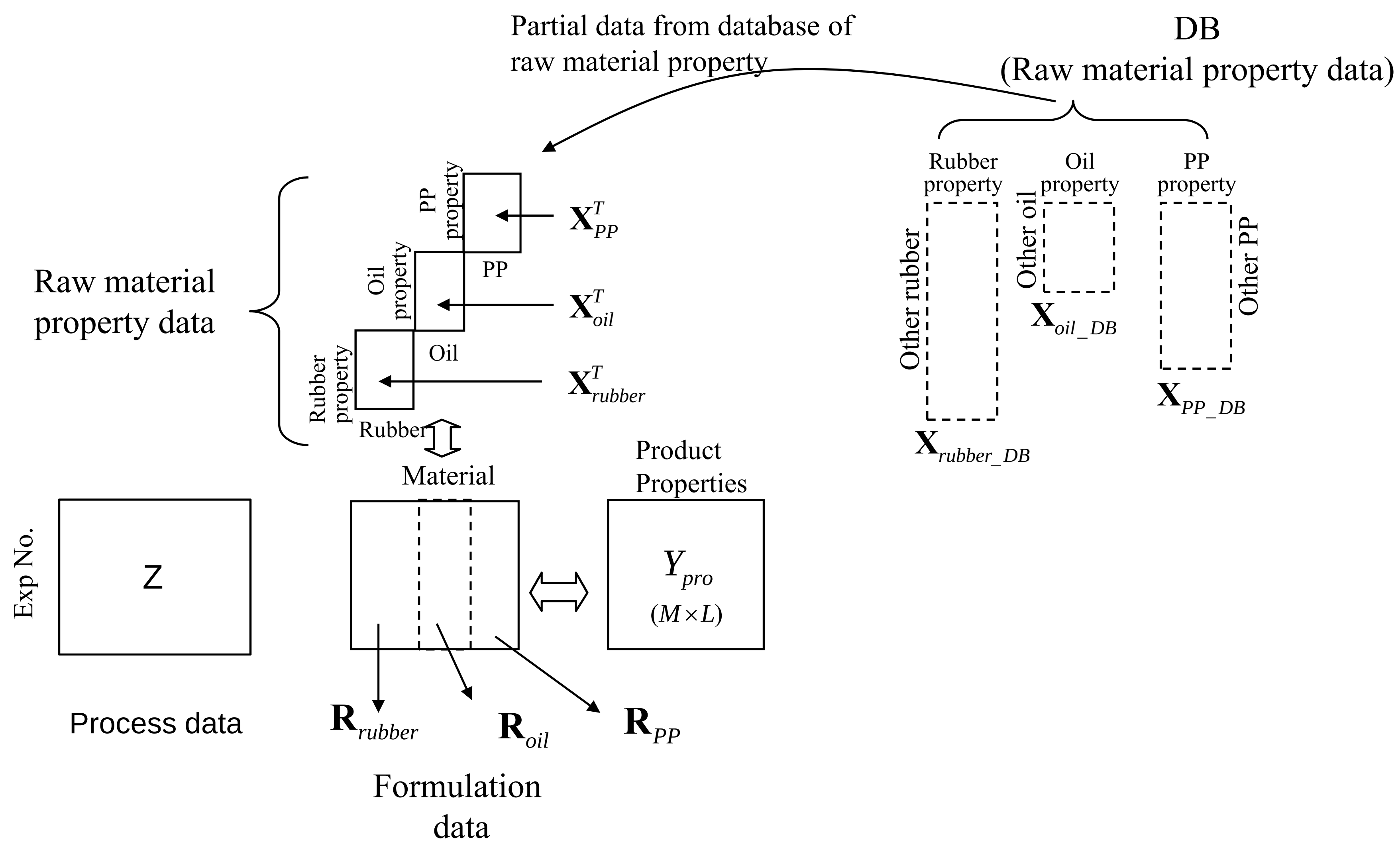
Example: Functional Polymer Development

Mitsubishi Chemicals



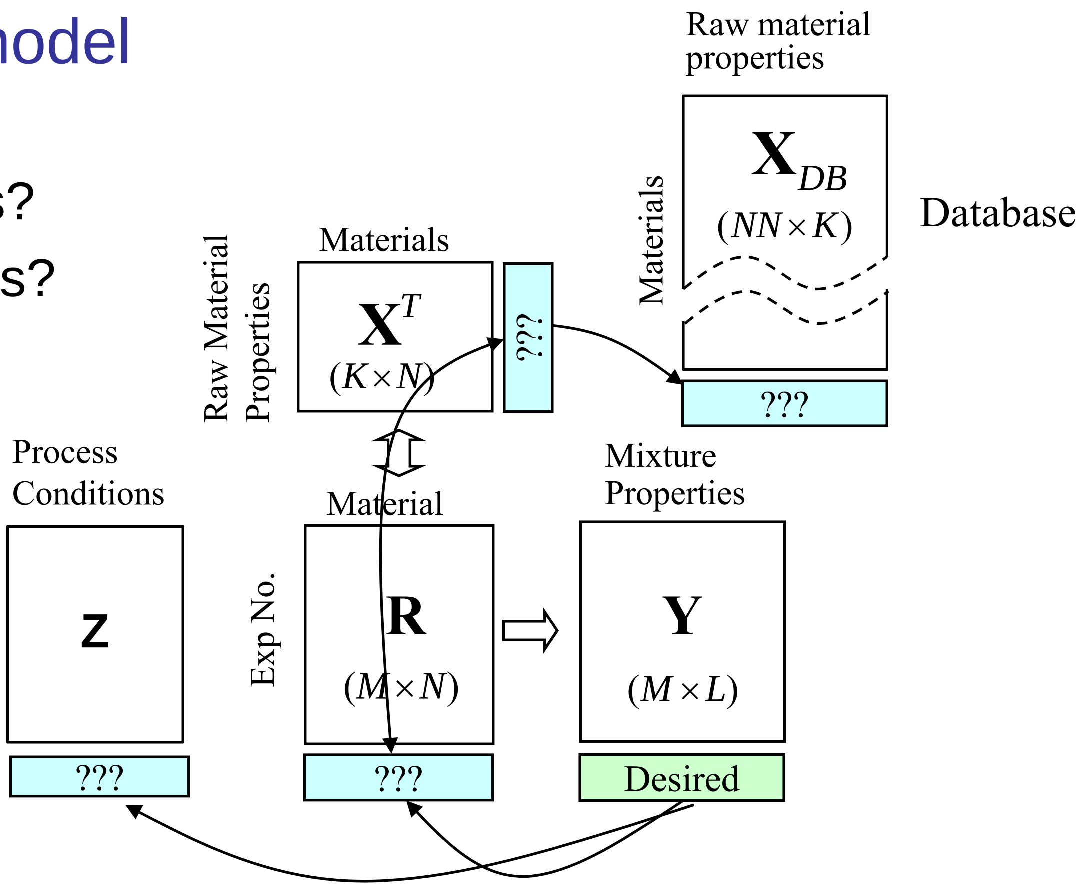
Equally applicable to pharmaceutical tablet formulations

Data structure



Methodology

- Build a multi-block PLS model that relates all the databases together and predicts the final quality attributes
- Perform an optimization in the latent variable space of the multi-block PLS model
  - Which materials?
  - Formulation ratios?
  - Process conditions?
  - Minimum cost



Formulation of the Optimization

$$\min_{r_{new}} \; \underbrace{(y_{des} - x_{mix\,new} \cdot B_{PLS})^T \cdot W_1 \cdot (y_{des} - x_{mix\,new} \cdot B_{PLS})}_{\text{Estimation error}} + w_2 \cdot \underbrace{\sum_{j=1}^{NN} r_{new,j} \cdot c_j}_{\text{Total material cost}} + w_3 \cdot \underbrace{\sum_{j=1}^{NN} \delta_j}_{\text{The number of materials}}$$

$s.t.$

Ideal mixing rule

$$x_{mix\,new} = r_{new} \cdot \mathbf{X}_{DB}$$

PLS model constraint

$$SPE_{new} = \sum_{k=1}^{k=K} (x_{mix\,new} - \hat{x}_{mix\,new})^2 \cong 0$$

$$T_{new}^{\;\;2} = \sum_{a=1}^{A} \frac{\tau_{new,a}^2}{s_a} \le const$$

Mixture constraint

$$\sum_{j=1}^{NN} r_{new,j} = 1, \;\; 0 \le r_{new,j} \le 1$$

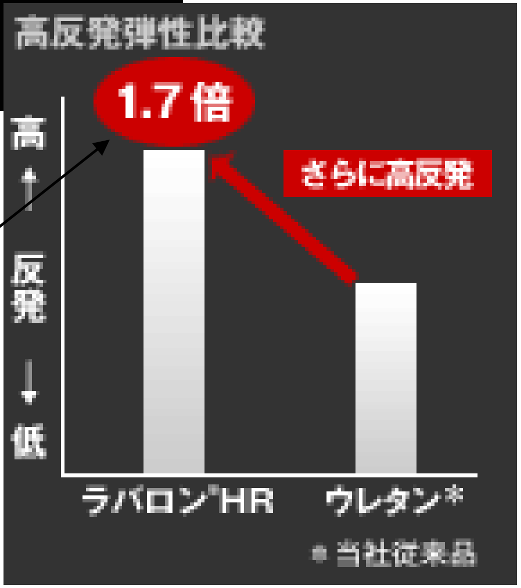
Binary variable constraint

$$\delta_j = \begin{cases} 1 & r_{new,j} > 0 \\ 0 & r_{new,j} = 0 \end{cases}$$

<u>Optimized variables</u>:
The mixture ratios of all the raw materials available on the database X_{DB} (and process variables Z)

Nonlinear, Constrained, Mixed Integer Optimization Problem

Example: Golf ball development



Approach to golf ball core design increased the resilience 1.7 times compared to previous products

C. LV approaches on industrial applications

- Analysis and monitoring of a batch process
- Control of final quality attributes in batch processes
- Optimization of process operation (batch)
- Scale-up and transfer between plants
- Rapid development of new products
- DOE in Latent Variable spaces to enrich dataset.

DOE's to enhance information content

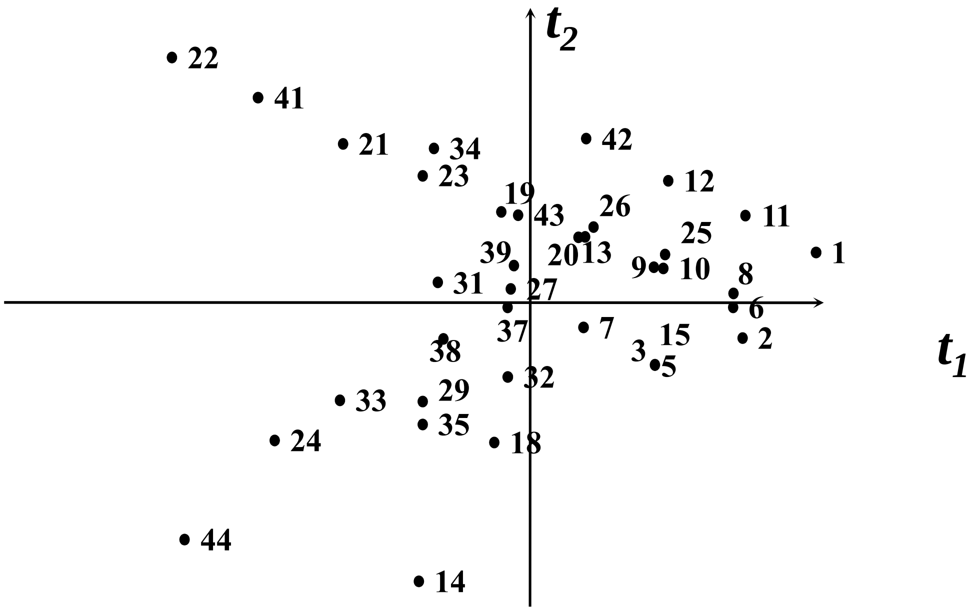
- Often industrial data bases are very large, but contain data only in limited regions
- Need add designed experiments to enhance information content of these large databases
  - But DOE space of original variables is extremely large!
  - DOE in LV space
- DOE's can be used to provide a small number of runs that can upgrade these databases
  - Example: for product development
  - DOE consists of simultaneous selection combinations of :
    - Raw materials
    - Formulation ratios
    - Processing conditions

  that will best enhance the information in the data-base

Concept of DOE in latent variable spaces

- Note regions of LV space where there are no data
- Use optimal DOE's to find those scores (t_1, t_2, ….) that would fill in these holes

Latent variable space

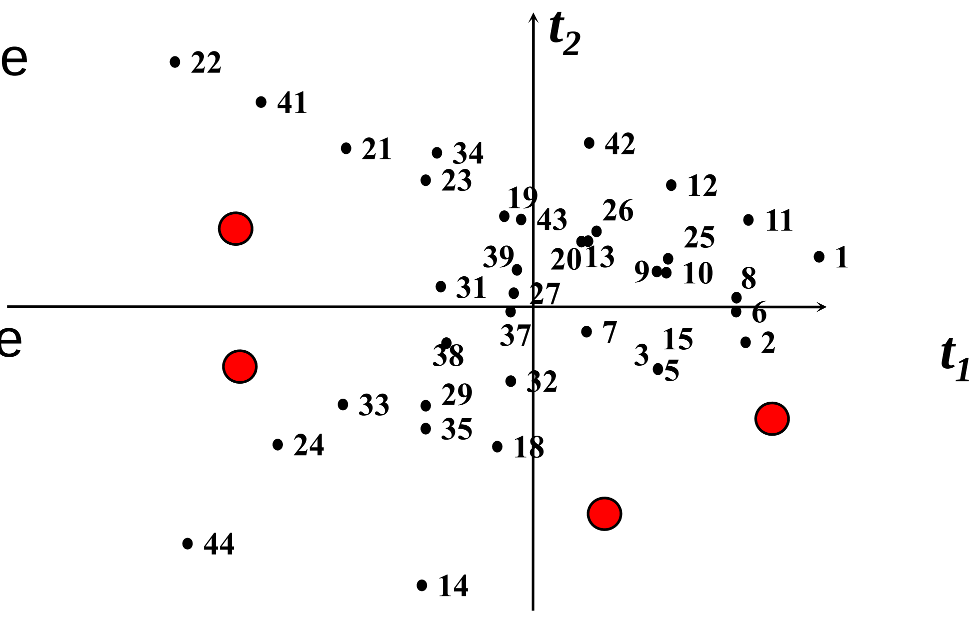


Muteki, K., J.F. MacGregor, and T. Ueda, "Mixture designs and models for the simultaneous selection of ingredients and their ratios", <u>Chemometrics & Intell. Lab. Systems</u>, 86, 17-25, 2007.

DOE in latent variable spaces

- Experiments (●) in score space
- From the DOE in the scores (t1, t2, …) use LV model of x-space to provide corresponding DOE in the raw materials, formulations and processing conditions: [Z, X, R]
- i.e. DOE in low dimensional score space provides a corresponding DOE in the high dimensional original variable space
- Very powerful concept
  - Drug design (SMD's)
  - Product development

Latent variable space



Summary

- Latent Variable methods for handling and integrating large volumes of industrial data
  - Concepts and motivation for latent variable methods
- Passive Applications:
  - Understanding through the analysis of historical data
  - On-line monitoring of process health
- Active Applications:
  - Control of final product quality
  - Optimizing process conditions
  - Scale-up and transfer between plants
  - Development of new products
  - DOE’s to enhance information content of the large databases


Thank You

Some References on topics in the presentation

- Latent variable methods (general)
  - Eriksson L., Johansson, E., Kettaneh-Wold, N. and Wold, S., 1999. “Introduction to Multi- and Megavariate Data Analysis using Projection Methods (PCA & PLS), Umetrics AB, Umea, Sweden
  - Kourti, T. (2002). Process Analysis and Abnormal Situation Detection: From Theory to Practice. IEEE Control Systems, 22(5), 10-25.
- Software
  - SIMCA_P (Umetrics); Unscrambler (Camo); Matlab toolbox (Eigenvector Technologies), ProMV (ProSensus)
- Analysis of historical data
  - Garcia-Munoz, S., T. Kourti and J.F. MacGregor, A.G.. Mateos and G. Murphy, “Trouble-shooting of an industrial batch process using multivariate methods”, Ind. & Eng. Chem. Res., 42, 3592-3601, 2003
- Monitoring
  - T. Kourti and J.F. MacGregor, 1995. "Process Analysis, Monitoring and Diagnosis Using Multivariate Projection Methods", J. Chemometrics and Intell. Lab. Systems, 28, 3-21.
- Control
  - Flores-Cerillo, J. and J. F. MacGregor, “Within-batch and batch-to-batch inferential adaptive control of semi-batch reactors: A Partial Least Squares approach”, Ind. & Eng. Chem. Res., 42, 3334-3345, 2003.
- Image-based soft sensors
  - Yu, H., J.F. MacGregor, G. Haarsma, and W. Bourg, “Digital imaging for on-line monitoring and control of industrial snack food processes”, Ind. & Eng. Chem. Res., 42, 3036-3044, 2003
  - Yu, H. and J.F. MacGregor, “Multivariate image analysis and regression for prediction of coating content and distribution in the production of snack foods”, Chem. & Intell. Lab. Syst., 67, 125-144, 2003

References, continued

- Optimization
  - Jaeckle, J.M., and MacGregor, J.F. (1998). Product Design Through Multivariate Statistical Analysis of Process Data. AIChE Journal, 44, 1105-1118.
  - Jaeckle, J.M., and MacGregor, J.F. (2000). Industrial Applications of Product Design through the Inversion of Latent Variable Models. Chemometrics and Intelligent Laboratory Systems, 50, 199-210.
  - Yacoub, F. and J.F. MacGregor, "Product optimization and control in the latent variable space of nonlinear PLS models", Chemometrics & Intell. Lab. Syst., 70, 63-74, 2004.
  - Garcia-Munoz, S., J.F. MacGregor, D. Neogi, B.E. Latshaw and S. Mehta, "Optimization of batch operating policies. Part II: Incorporating process constraints and industrial applications", Ind. & Eng. Chem. Res., Published on-line, May, 2008
- Product development
  - Muteki, K., J.F. MacGregor and T. Ueda, "On the Rapid development of New Polymer Blends: The optimal selection of materials and blend ratios", Ind. & Eng. Chem. Res., 45, 4653-4660, 2006.
  - Muteki, K. and J.F. MacGregor, "Multi-block PLS Modeling for L-shaped Data Structures, with Applications to Mixture Modeling", Chemometrics & Intell. Lab. Systems, 85, 186-194, 2006
- Design of Experiments
  - Muteki, K., J.F. MacGregor, and T. Ueda, "Mixture designs and models for the simultaneous selection of ingredients and their ratios", Chemometrics & Intell. Lab. Systems, 86, 17-25, 2007.
  - Muteki, K. and J.F. MacGregor, "Sequential design of mixture experiments for the development of new products", Chemometrics & Intell. Lab Sys., 2007.