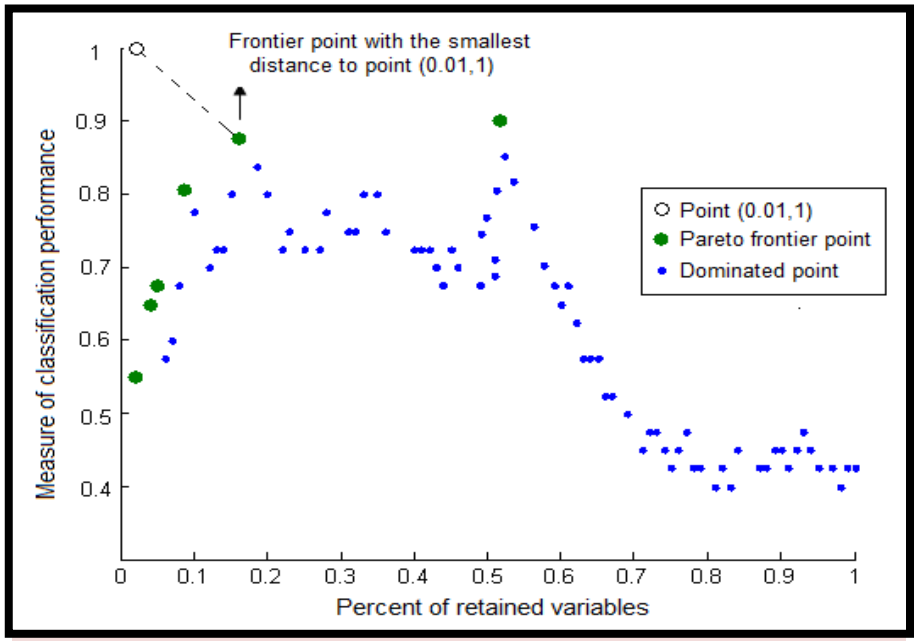
Data Mining: Feature Selection for Batch Production

Susan Albin, Art Chaovalitwongse, Michel Anzanello

- Datasets containing a large number of noisy and correlated process variables are commonly found in chemical and industrial processes, making it hard for engineers to identify the key features.
- The focus is on correctly classifying the outcome of each production batch based on rpoduction variables or features.
- The objective is to reduce number of process features for classification by eliminating noisy and irrelevant ones.

Dataset	Number of Process Variables	Number of observations	
		Training set	Testing set
ADPN	100	57	14
GRANU	78	300	200
LATEX	117	210	52
OXY	95	300	200
PAPER	54	192	192
SPIRA	96	115	29

Testing proposed feature selection on 6 real data sets.



It is optimal to select only 10% of available features to correctly classify production batches in LATEX production.

	Pareto Variable Selection (PVS)	Stepwise Multiple Regression (SMR)	Simple Regression (SR)
Classification Accuracy (%)	77	74	74
Classification Accuracy Standard Deviation (%)	3	7	7
Retained Variables (%)	6	25	39
Retained Variables Standard Deviation (%)	3	6	6

Performance of PVS and traditional methods for variable selection on testing sets of simulated data

$$v_j = \sum_{a=1}^{A} (w_{ja}^*)^2 R_{Ya}^2$$

importance index of process variable *j* derived from PLS weights and loadings