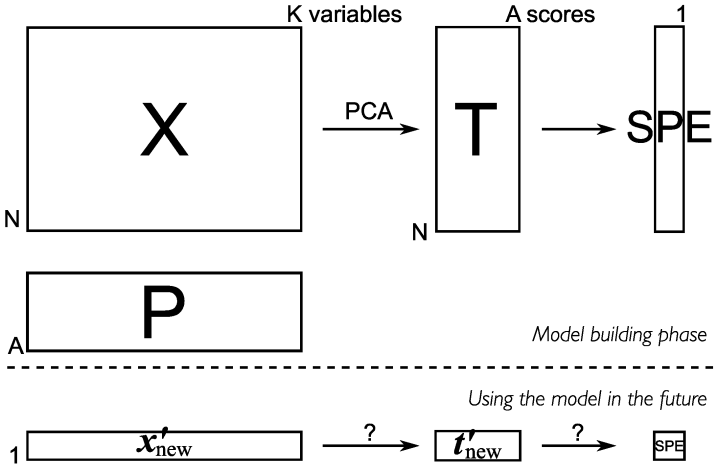
Multivariate Statistical Methods for Big Data Analysis and Process Improvement

Lecture 5: Using PCA on new data - Process Monitoring

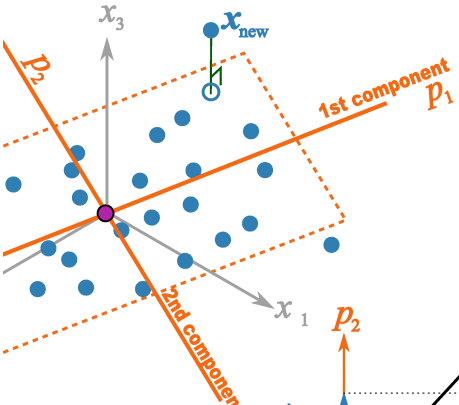
Dr. Brandon Corbett

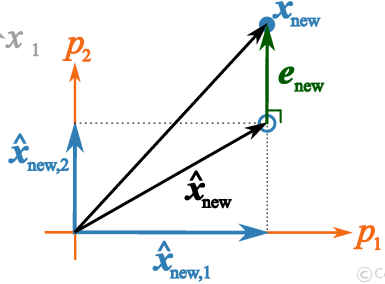
Course notes for ChE 765/SEP 767, McMaster University

Using a PCA model: concept



Using a PCA model: geometric concept





$\mathbf{t}'_{new} =$
$\hat{\mathbf{x}}'_{new} =$
$\mathbf{e}'_{new} =$
$SPE_{new} =$

Using existing PCA model on new data

Very simple:

- Preprocess the raw data: $\mathbf{x}_{\text{new,raw}} \longrightarrow \mathbf{x}_{\text{new}}$
- Project onto existing model to get scores: $\mathbf{t}'_{\text{new}} = \mathbf{x}'_{\text{new}}\mathbf{P}$
- Calculated predicted $\hat{\mathbf{x}}'_{\text{new}} = \mathbf{t}'_{\text{new}}\mathbf{P}'$
- Calculate residuals: $\mathbf{e}'_{\text{new}} = \mathbf{x}'_{\text{new}} - \hat{\mathbf{x}}'_{\text{new}}$
- Calculate $\text{SPE}_{\text{new}} = \mathbf{e}'_{\text{new}}\mathbf{e}_{\text{new}} = \text{ssq}(\mathbf{e}_{\text{new}})$

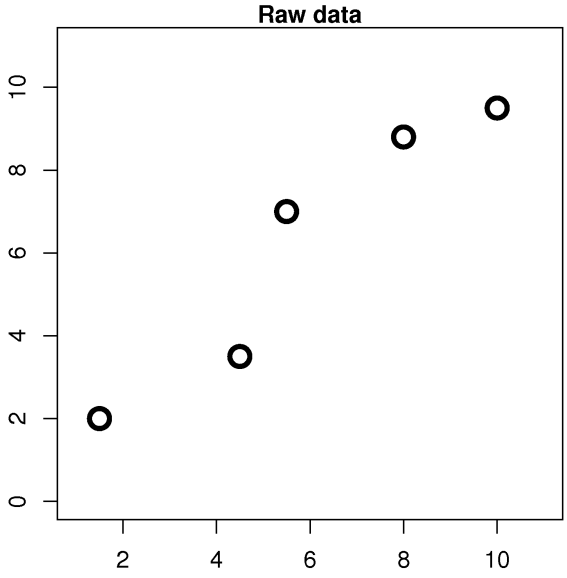
General problem: overfitting

Overfitting

Adding complexity to model when it's not supported **by the data**

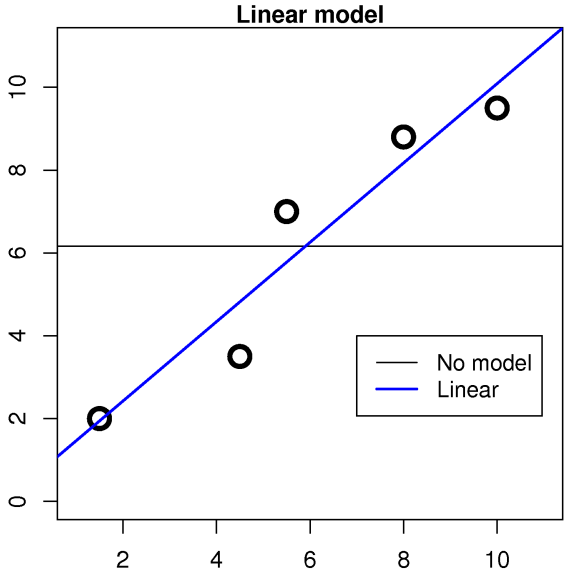
Overfitting is a problem that can occur with *any model*

Overfitting example



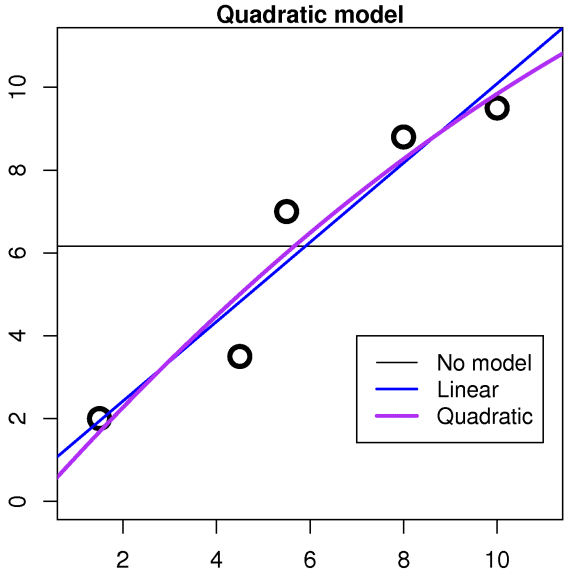
What model would you fit initially?

Overfitting example



$R^2_{\text{linear}} = 0.908$

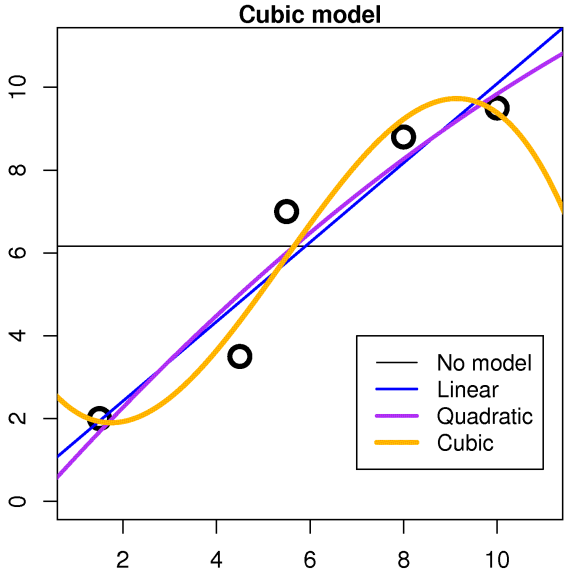
Overfitting example



$R^2_{\text{linear}} = 0.908$

$R^2_{\text{quad}} = 0.913$

Overfitting example

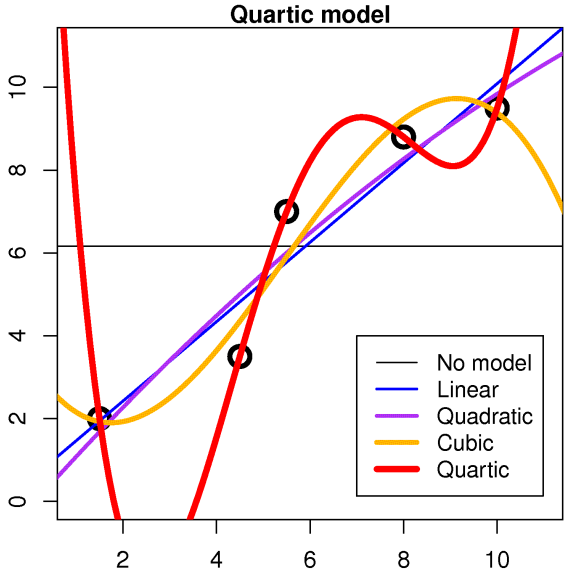


$R^2_{\text{linear}} = 0.908$

$R^2_{\text{quad}} = 0.913$

$R^2_{\text{cubic}} = 0.951$

Overfitting example



$R^2_{\text{linear}} = 0.908$

$R^2_{\text{quad}} = 0.913$

$R^2_{\text{cubic}} = 0.951$

$R^2_{\text{quartic}} = 1.00$

Overfitting example

- **What's the problem here?**

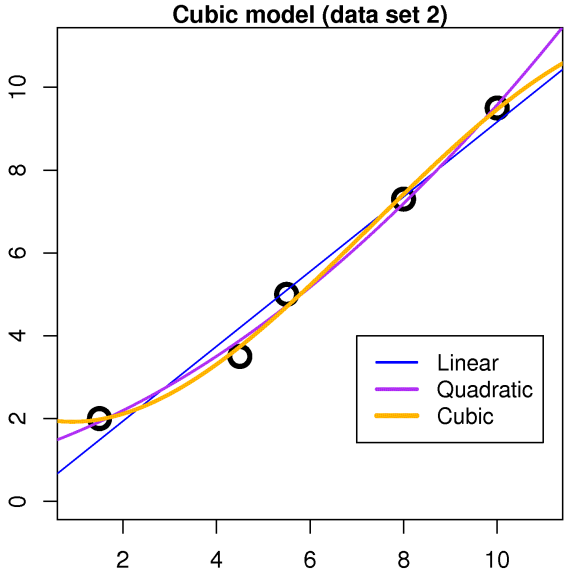
Overfitting issue

Knowing how the model will be used in the future is the key to avoiding overfitting

How many terms would you use in a model if you were:

- describing the process to a colleague
- making predictions of y given a new, unknown x
- making predictions of x for a desired y

Overfitting example 2



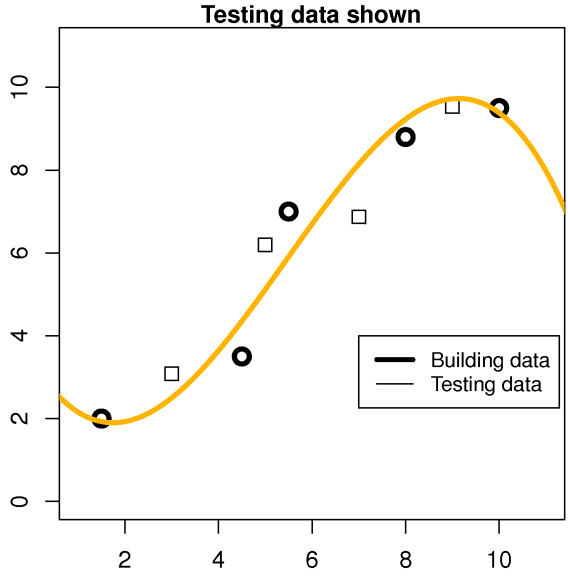
Still agree of your previous answers?

Avoid overfitting: use a testing data set

Principle:

1. Keep a testing data set aside
2. Predict all values from testing data; e.g.
   - $\hat{y}_i = b_0 + b_1 x_i + b_2 x_i^2$
3. calculate the residuals
   - $e_i = y_i - \hat{y}_i$
4. Calculate $\text{ssq}(e_i)$ = *prediction error sum of squares* = PRESS

Example: using the testing set

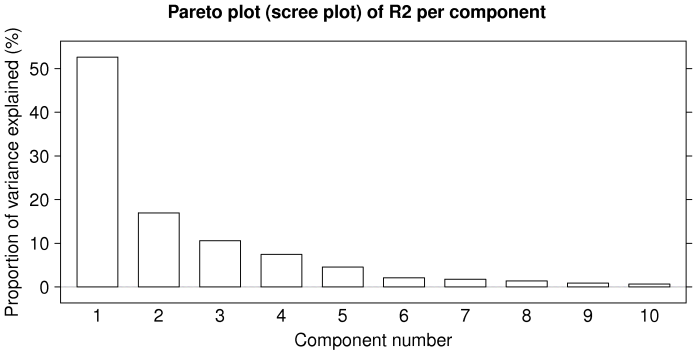


Number of model parameters and PRESS:

1. 21.4 (no model)
2. 1.17 (linear)
3. 1.05 (quadratic)
4. 3.15 (cubic)
5. 29.3 (quartic)

The number of components to use: the problem

- PCA's objective is to best explain data
- Over-fitting is when we add more components than are supported by the data in **X**



How not to do it ...

Fitting to achieve a certain R^2

Should not do this, but it is common to see this approach.

The number of components

Ideal approach

1. Keep a testing data set aside
2. Fit a component to training data
3. Project testing data onto model
4. Calculate the prediction error sum of squares (PRESS) on testing data
5. Repeat from step 2

- What should happen to PRESS as A increases?
- What if we don't have enough data to keep aside?

Cross-validation for PCA

- A *general* tool; can be applied to any model
- Cross-validation can help avoid over-fitting
- $\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E}_A = \mathbf{t}_1\mathbf{p}_1' + \mathbf{t}_2\mathbf{p}_2' + \ldots + \mathbf{t}_A\mathbf{p}_A' + \mathbf{E}_A$
- $\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E}_A$

Cross-validation's aim

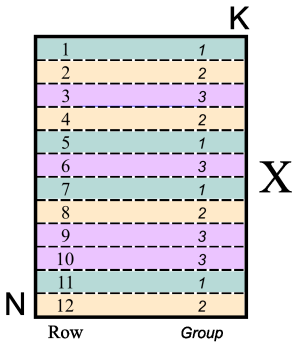
Find when residuals in $\mathbf{E}_A$ are "*small enough*" so there is no more information left

Cross-validation for PCA

- $\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E}_A$
- $\mathbf{X} = \widehat{\mathbf{X}} + \mathbf{E}_A$
- $\mathcal{V}(\mathbf{X}) = \mathcal{V}(\widehat{\mathbf{X}}) + \mathcal{V}(\mathbf{E}_A)$
- Recall: $R^2 = 1 - \dfrac{\mathcal{V}(\mathbf{E}_A)}{\mathcal{V}(\mathbf{X})}$
- Define: $Q^2 = 1 - \dfrac{\mathcal{V}(\text{predicted } \mathbf{E}_A)}{\mathcal{V}(\mathbf{X})}$
- $\mathcal{V}(\text{predicted } \mathbf{E}_A) = \text{PRESS} = \text{prediction error sum of squares}$

Cross-validation concept for PCA

Split the rows in **X** into G groups.

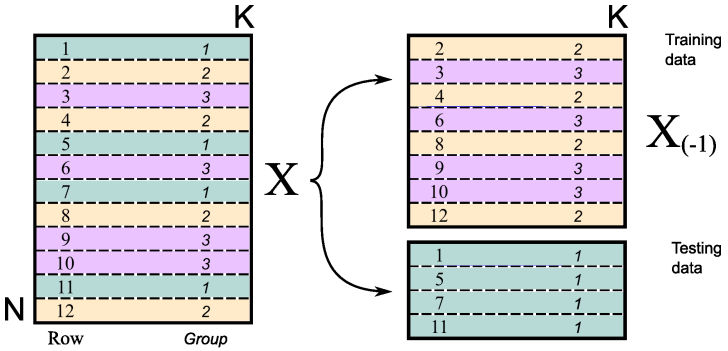


- Typically $G \approx 7$ [ProSensus, Simca-P use $G = 7$]
- Rows can be randomly grouped, or
- ordered *e.g.* 1, 2, 3, 1, 2, 3, ...

$G = 3$ in this illustration

Cross-validation concept for PCA

Fit a PCA model using $\mathbf{X}_{(-1)}$; use $\mathbf{X}_{(1)}$ as testing data

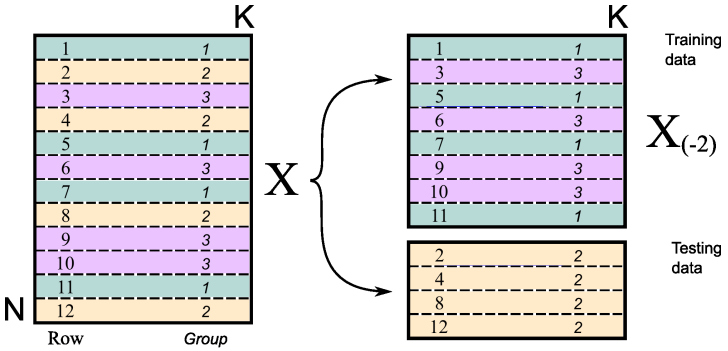


$\mathbf{E}_{(1)}$ = prediction error for testing group 1

 26

Cross-validation concept for PCA

Fit a PCA model using $\mathbf{X}_{(-2)}$; use $\mathbf{X}_{(2)}$ as testing data

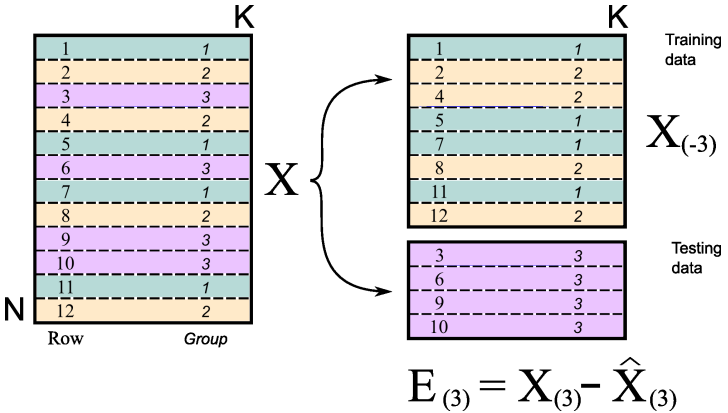


$\mathbf{E}_{(2)}$ = prediction error for testing group 2

 27

Cross-validation concept for PCA

Fit a PCA model using $\mathbf{X}_{(-3)}$; use $\mathbf{X}_{(3)}$ as testing data



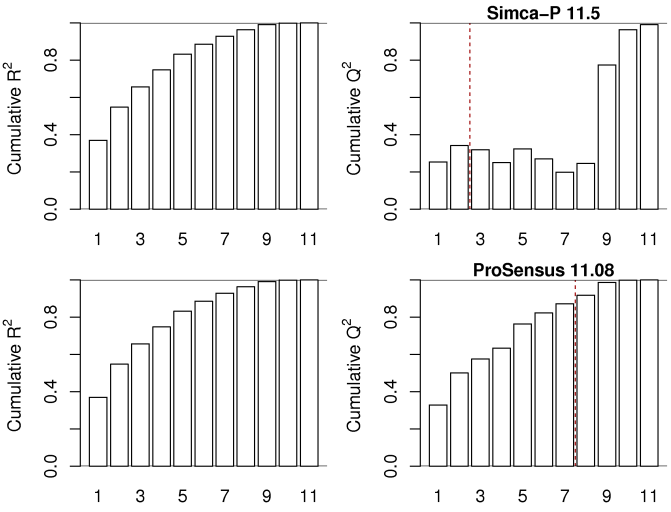
$\mathbf{E}_{(3)}$ = prediction error for testing group 3

 28

Q^2 calculation and interpretation

- PRESS $= \text{ssq}(\mathbf{E}_{(1)}) + \text{ssq}(\mathbf{E}_{(2)}) + \ldots + \text{ssq}(\mathbf{E}_{(G)})$
- PRESS $=$ prediction error sum of squares from each prediction group
- $Q^2 = 1 - \dfrac{\mathcal{V}(\text{predicted } \mathbf{E}_A)}{\mathcal{V}(\mathbf{X})} = 1 - \dfrac{\text{PRESS}}{\mathcal{V}(\mathbf{X})}$
- Q^2 is calculated and interpreted in the same way as R^2
- Q_k^2 can be calculated for variable $k = 1, 2, \ldots K$
- You should always find $Q^2 \leq R^2$
- If $Q^2 \approx R^2$: that component is useful and predictive in the model
- If Q^2 is "small": that component is likely fitting noise

Cross-validation: Q^2 can differ in packages



Using default settings in both packages ($G = 7$)

 30

Cross-validation and "autofit"

Each package differs. General idea is to keep the component if

- Model's Q^2 increases by some percentage (e.g 1%)
- Any variable's Q^2 increases by some amount (e.g. 5%)

Summary: number of components to use

- Cross-validation (autofit) is an "OK" guide: but always test
- Still an open topic (no single correct method)
  - Resampling methods are a current research topic
- Always fit a few extra components in software
- Do the extra PCs mean anything? Do they help solve your objective? **If so: keep them**

Monitoring analogy: your health

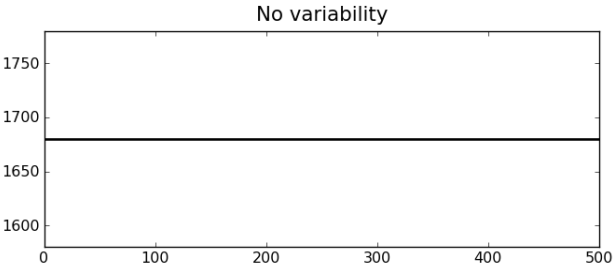
- You have an intuitive (built-in) model for your body
- When everything is normal: we say "*I'm healthy*" (in control)
- **Detect a problem**: pain, lack of mobility, hard to breath
- Something feels wrong (there's a special cause)
- **Diagnose the problem**: yourself, search internet, doctor
- Fix the problem and get back to your usual healthy state

Monitoring analogy: your health

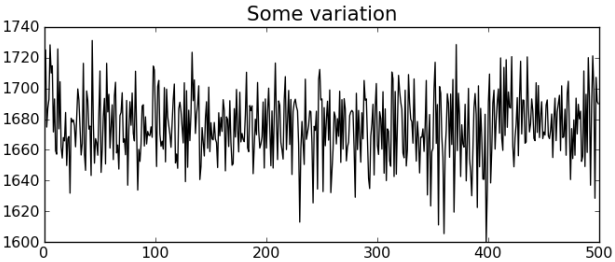
Where did that intuitive model for your body's health come from?

Monitoring concept for a process

Our goal: We want process stability



Variability

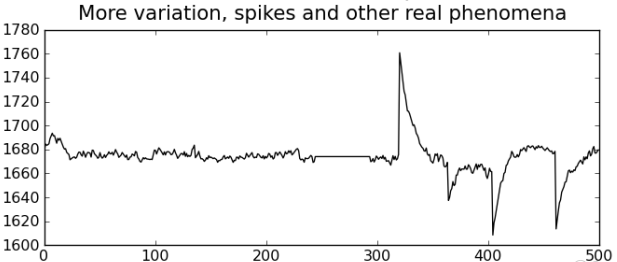


Best case: we have **unaccounted sources** of noise: called **error**

Variability

More realistically:

- Sensor drift, spikes, noise, recalibration shifts, errors in our sample analysis
- Operating staff: introduce variability into a process
- Raw material properties are not constant
- External conditions change (ambient temperature, humidity)
- Equipment breaks down, wears out, sensor drift, maintenance shut downs
- Feedback control introduces variability



 46

Variability in your product

Assertion

Customers expect both uniformity and low cost when they buy your product. Variability defeats both objectives.

Remind yourself of the last time you bought something that didn't work properly

So what do we want

1. *rapid* problem detection
2. diagnose the problem
3. finally, adjust the process so problems don't occur

Process monitoring is mostly **reactive** and not *proactive*. So it is suited to *incremental* process improvement

Process Monitoring: Terminology

- "Process monitoring" is also called "Statistical Process Control" (SPC)
- We will avoid this term to avoid confusion with feedback control

Other buzzwords

- Dashboards
- Analytics
- BI: Business intelligence
- KPI: Key performance indicators

Note that these tools are not limited to manufacturing, also applied in finance, markets, retail, and others...

Monitoring analogy: making errors

Assume the doctor is always right and that the baseline hypothesis is: "*you are healthy*"

- **Type 1 error**: *you detect a problem* (e.g. hard to breathe); doctor says nothing is wrong
  - You've raised a false alarm
  - You feel outside your limits,
  - but the truth is: "*you are healthy*"
  - **Type 1 error** = raise an alarm when there isn't a problem

Monitoring analogy: making errors

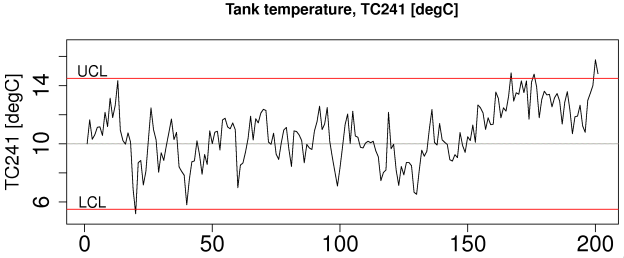
Assume the doctor is always right and that the baseline hypothesis is: "*you are healthy*"

- **Type 2 error**: *you feel OK*; but go to doctors for physical and they detect a problem
  - You feel within your limits,
  - but the truth is: "*you are not healthy*"
  - **Type 2 error** = don't raise an alarm when there is a problem
- The grid

Shewhart chart (recap)

- Named for *Walter Shewhart* from Bell Telephone and Western Electric, parts manufacturing, 1920's
- A chart for monitoring variable's *location*, shown with
- a lower control limit (LCL), usually at $+3\sigma$
- a upper control limit (UCL), usually at -3σ
- a target, at the setpoint/desired value

No action taken as long as the variable plotted remains within limits (in-control). Why?



Judging the chart's performance

- **Type I error**:
  - value plotted is from common-cause operation, but falls outside limits
  - if values are normally distributed, how many will fall outside?
    - $\pm 2\sigma$ limits?
    - $\pm 3\sigma$ limits?
  - *Synonyms*: false alarm, producer's risk
- **Type II error**:
  - value plotted is from abnormal operation, but falls inside limits
  - *Synonyms*: false negative, consumer's risk

Adjusting the chart's performance

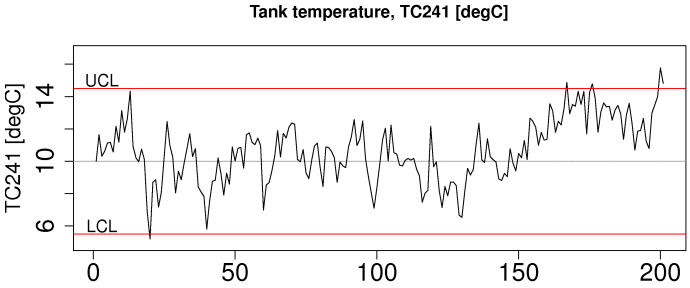
Key point

Control chart limits are not set in stone. Adjust them!

Nothing makes a control chart more useless to operators than frequent false alarms.

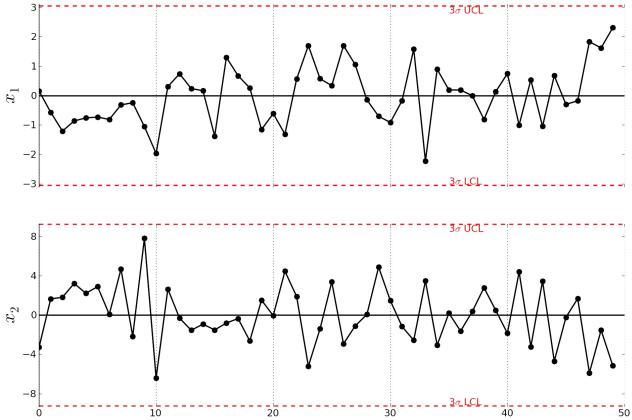
- **But, you cannot simultaneously have low type I and type II error**

Discussion



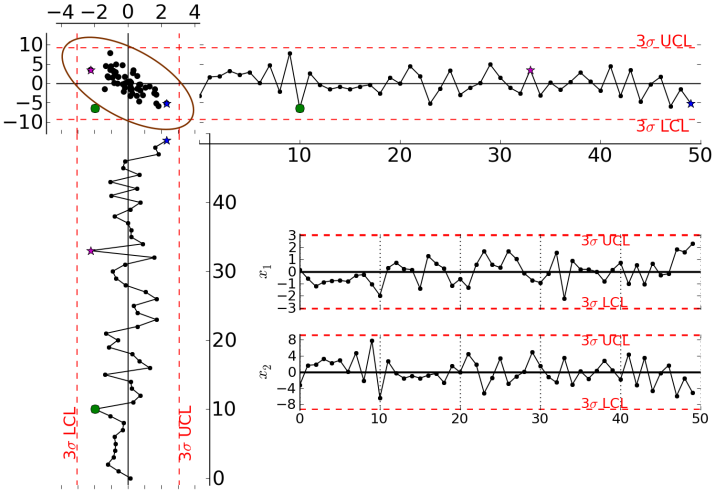
1. What action is taken when outside the limits
2. What if data goes missing?

Discussion



3. Monitoring many variables.
   - Feasible?
   - Is each plot showing something new?

Discussion: multivariate monitoring



Process monitoring with PCA: scores

Monitoring with latent variables; use:

- scores from the model, $t_1, t_2, \ldots, t_A$

Illustration on the board

Limits for other PCA model quantities: scores

1. Use a large historical data set (*shown on board*)
2. Use a statistical distribution assumption
   - $t_a \sim \mathcal{N}(0, s_a)$ (why?)
   - $100(1-\alpha)\%\text{limit} = \pm \left(t_{\alpha/2,df}\right) s_a$
   - s_a = standard deviation of score column a
   - df = degrees of freedom $= N - 1$

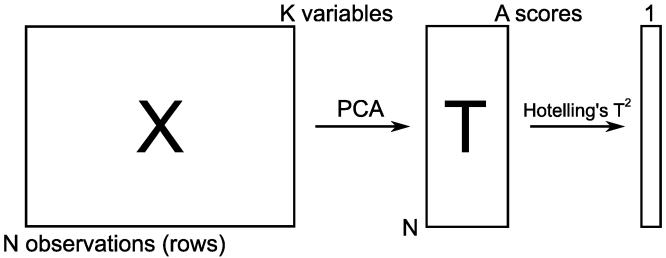
Process monitoring with PCA: scores

Much better than the raw variables:

- The scores are orthogonal (independent)
- Far fewer scores than original variables
- Calculated even if there are missing data
- Can be monitored anywhere there is real-time data
- Available before the lab's final measurement

Hotelling's T^2

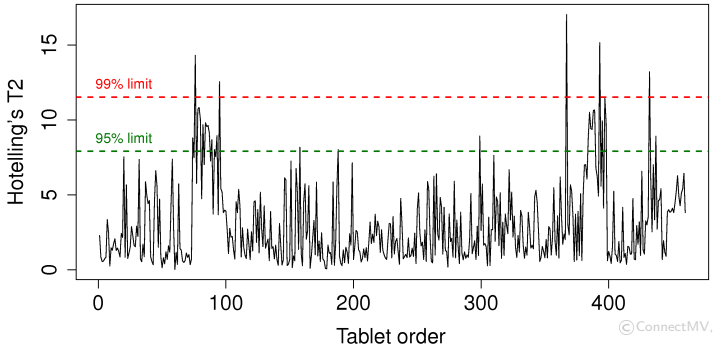
- After extracting components from **X** we accumulate A score vectors in matrix **T**



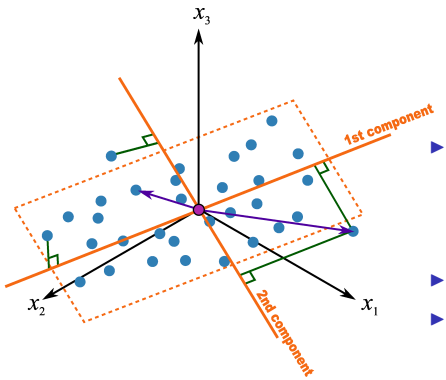
- T_i^2 is a summary of all A components within row i
- $T_i^2 = \sum_{a=1}^{a=A} \left(\frac{t_{i,a}}{s_a} \right)^2$
- s_a = standard deviation of score column a

Hotelling's T^2

- $T_i^2 = \sum_{a=1}^{a=A} \left(\frac{t_{i,a}}{s_a} \right)^2$
  - $s_1 > s_2 > \ldots$ (from the eigenvalue derivation)
  - $T_i^2 \geq 0$
  - Plotted as a time-series/sequence plot
  - Useful if the row order in dataset has a meaning



Hotelling's T^2



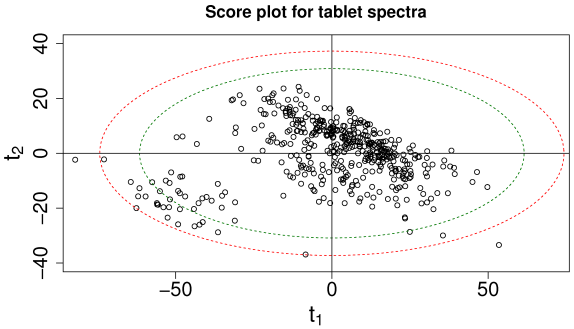
$$T_i^2 = \sum_{a=1}^{a=A} \left(\frac{t_{i,a}}{s_a} \right)^2 \geq 0$$

- Interpretation: directed distance from the center to where the point is projected on the plane
- T^2 has an F-distribution
- Often show the 95% confidence limit value, called $T^2_{A,\alpha=0.05}$

 35

Hotelling's T^2

- If $A = 2$, equation for 95% limit $= T^2_{A=2,\alpha=0.05} = \frac{t_1^2}{s_1^2} + \frac{t_2^2}{s_2^2}$
- An equation for an ellipse
- s_1 and s_2 are constant for a given model
- Points on ellipse have a constant distance from model center



Hotelling's T^2

$$T^2_{A,\alpha} = \frac{(N-1)(N+1)A}{N(N-A)} \cdot F_\alpha(A, N-A)$$

- for A components
- on N observations
- for the $100(1-\alpha)\%$ confidence limit

T^2 is also known as:

- Mahalanobis distance (see paper on literature website)
- D-statistic

Limits for other PCA model quantities: SPE

1. Use a large historical data set (*shown on board*)
2. $\text{SPE}_i \sim g\chi^2(h)$ (approximately)
   - $g = \dfrac{v}{2m}$ = premultiplier
   - $h = \dfrac{2m^2}{v}$ = degrees of freedom of $\chi^2(h)$
   - $m = \text{mean}(SPE)$
   - $v = \text{var}(SPE)$
3. We use the data to estimate g and h
4. See derivation in Nomikos and MacGregor - paper 34

Process monitoring with PCA: Hotelling's T^2

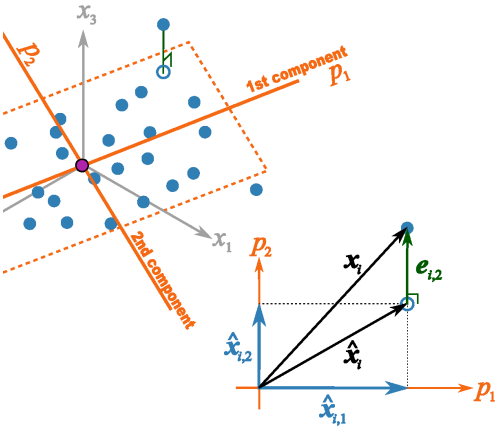
Hotelling's $T^2 = \sum_{a=1}^{a=A} \left(\frac{t_a}{s_a} \right)^2$

- The distance along the model plane
- Is a one-side monitoring plot
- What does a large T^2 value mean?

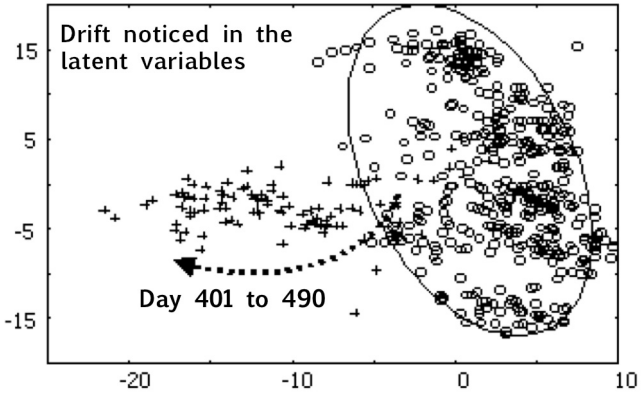
Process monitoring with PCA: SPE

$\text{SPE}_i = (\mathbf{x}_i - \hat{\mathbf{x}}_i)'(\mathbf{x}_i - \hat{\mathbf{x}}_i) = \mathbf{e}_i'\mathbf{e}_i$

- Distance off the model plane
- Is a one-side monitoring plot
- What does a large SPE value mean?



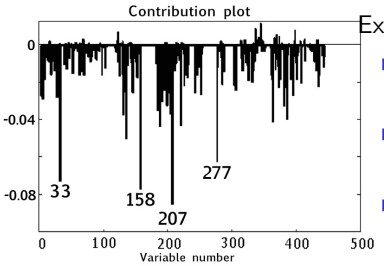
Diagnosing a problem



- Interrogate the latent variables to see what changed

LVM for troubleshooting: contribution plot

- Shows difference between two points in the score plot



Example:

- **207**: temperature on tray 129 in distillation column 3
- **158**: a tag from distillation column 3
- **33** and **277**: related to concentration of feed A

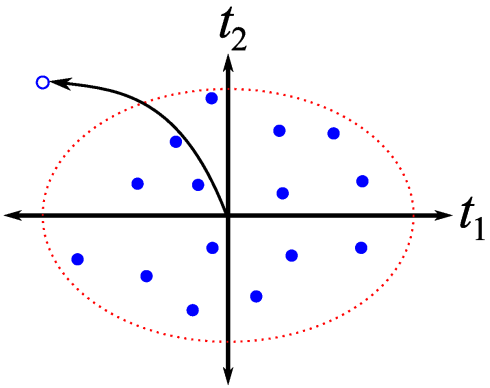
- These variables are related to the problem
- *Not the cause of the problem*
- Still have to use your engineering judgement to diagnose
- But, we've reduced the size of the problem

Contribution plots

- Scores: $t_{i,a} = \mathbf{x}_i \mathbf{p}_a$
  - $\begin{bmatrix} x_{i,1}p_{1,a} & x_{i,2}p_{2,a} & \dots & x_{i,k}p_{k,a} & \dots & x_{i,K}p_{K,a} \end{bmatrix}$
  - *Derivation on the board*

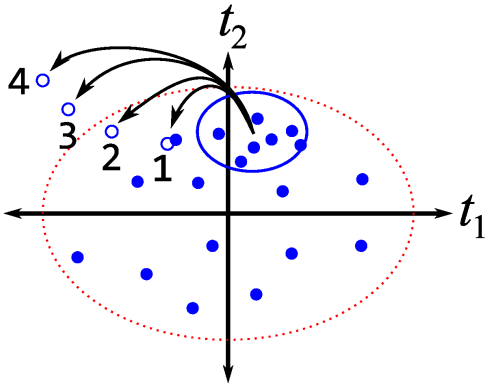
- T^2 contributions: weighted sum of scores
  - More details in Alvarez *et al.* - paper 21
  - and Kourti and MacGregor - paper 81

Contributions in the score space



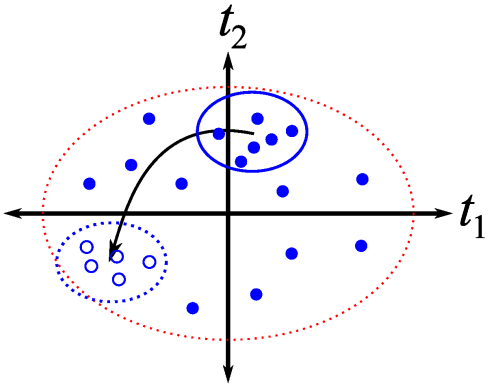
From the model center to a point

Contributions in the score space



Four seperate contribution plots to learn why the sequence of deviations occurred

Contributions in the score space



From one group to another group

Contribution plots

- SPE $= \mathbf{e}_i' \mathbf{e}_i$
  - where $\mathbf{e}_i' = \mathbf{x}_i' - \widehat{\mathbf{x}}_i'$
  - $\begin{bmatrix}(x_{i,1} - \hat{x}_{i,1}) & (x_{i,2} - \hat{x}_{i,2}) & \dots & (x_{i,K} - \hat{x}_{i,K})\end{bmatrix}$

- Joint T^2 and SPE monitoring plots
  - *Illustrated on the board*
  - Discussion

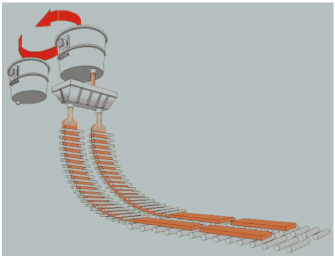
Industrial case study: Dofasco

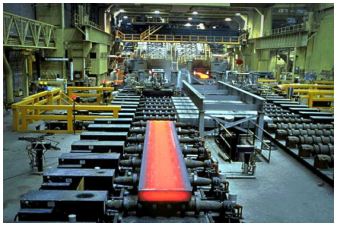
- ArcelorMittal in Hamilton (formerly called Dofasco) has used multivariate process monitoring tools since 1990's
- Over 100 applications used daily
- Most well known is their casting monitoring application, Caster SOS (Stable Operation Supervisor)
- It is a multivariate monitoring system

Dofasco case study: slabs of steel

All screenshots with permission of Dr. John MacGregor

Dofasco case study: casting

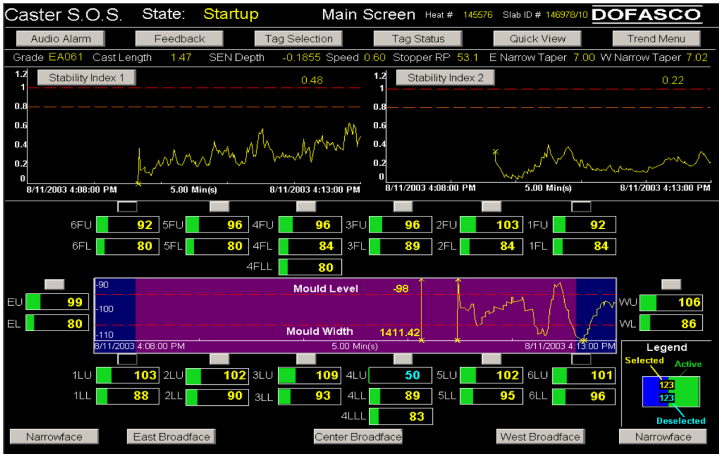




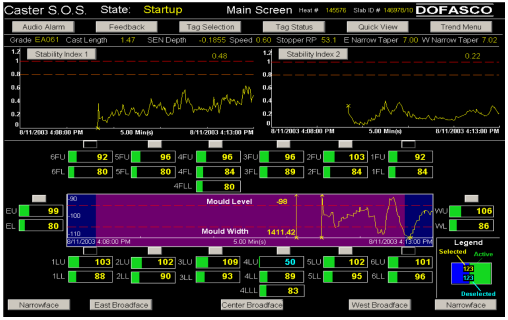
Dofasco case study: breakout

©ConnectMV, 2011 75

Dofasco case study: monitoring for breakouts



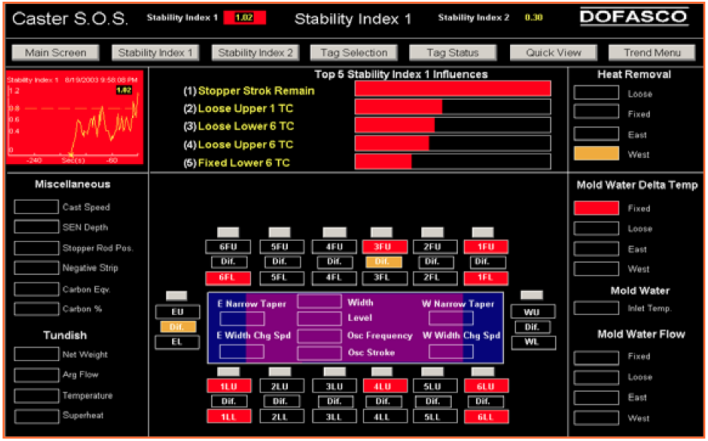
Dofasco case study: monitoring for breakouts



- Stability Index 1 and 2: one-sided monitoring chart
- Warning limits and the action limits.
- A two-sided chart in the middle
- Lots of other operator-relevant information

 77

Dofasco case study: an alarm



General procedure to build monitoring models I

1. Identify variable(s) to monitor.
2. Retrieve historical data (computer systems, or lab data, or paper records)
3. Import data and just plot it.
   - Any time trends, outliers, spikes, missing data gaps?
4. Locate regions of stable, common-cause operation.
   - Remove spikes and outliers
5. Building monitoring model
6. Model includes control limits (UCL, LCL) for scores, SPE and Hotelling's T^2
7. Test your chart on **new, unused** data.
   - Testing data: should contain both common and special cause operation
8. How does your chart work?
   - Quantify the type I and II error.

General procedure to build monitoring models II

- Adjust the limits;
- Repeat this step, as needed to achieve levels of error

9. Run chart on your desktop computer for a couple of days
   - Confirm unusual events with operators; would they have reacted to it? False alarm?
   - Refine your limits
10. Not an expert system - will not diagnose problems:
    - use your engineering judgement; look at patterns; knowledge of other process events
11. Demonstrate to your colleagues and manager
    - But go with dollar values
12. Installation and operator training will take time
13. Listen to your operators
    - make plots interactive - click on unusual point, it drills-down to give more context