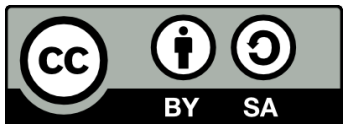
Latent Variable Methods Course

Learning from data

Instructor: Kevin Dunn
kevin.dunn@connectmv.com
http://connectmv.com

Revision: 268:adfd compiled on 15-12-2011

Copyright, sharing, and attribution notice



This license allows you:

- **to share** - to copy, distribute and transmit the work
- **to adapt** - but you must distribute the new result under the same or similar license to this one
- **commercialize** - you *are allowed* to create commercial applications based on this work
- **attribution** - you must attribute the work as follows:
  - "Portions of this work are the copyright of ConnectMV", *or*
  - "This work is the copyright of ConnectMV"

We appreciate:

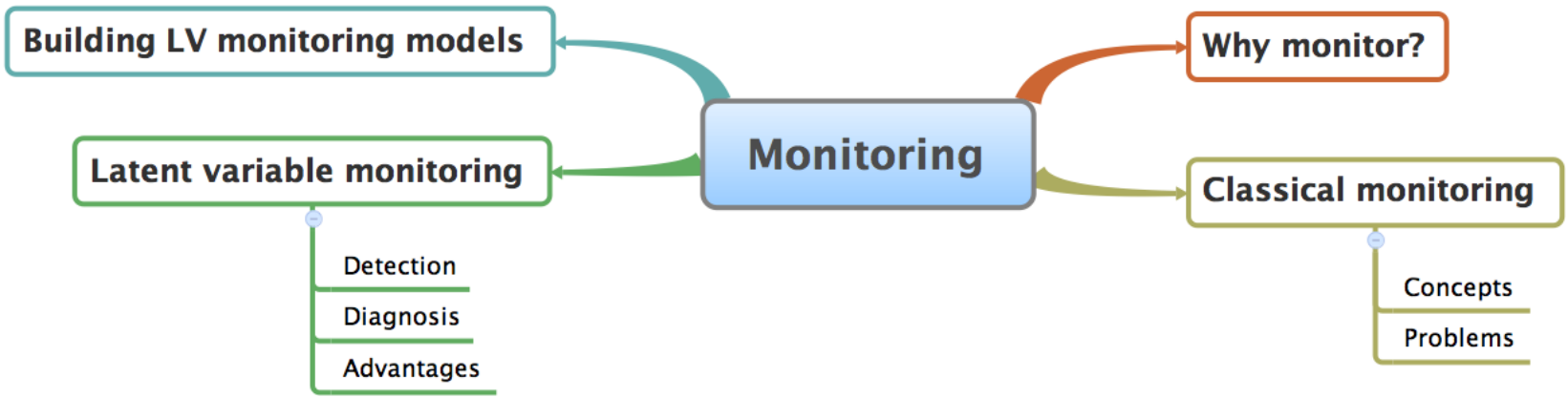
- if you let us know about **any errors** in the slides
- **any suggestions to improve the notes**
- telling us if you use the slides, especially commercially, so we can inform you of major updates
- emailing us to ask about different licensing terms

All of the above can be done by writing us at

`courses@connectmv.com`

If reporting errors/updates, please quote the current revision number: `268:adfd`

Summary of Process Monitoring



Review of assignment

How do I know a point is an outlier?

- Easier if it's your own data
- Which plots should I use to detect outliers?
- What a 95% limit means ...
  - Always confirm your conclusions from the raw data
  - Still have to use your head!

Activating the software

- Please email your codes to: *academic.promv@prosensus.ca*

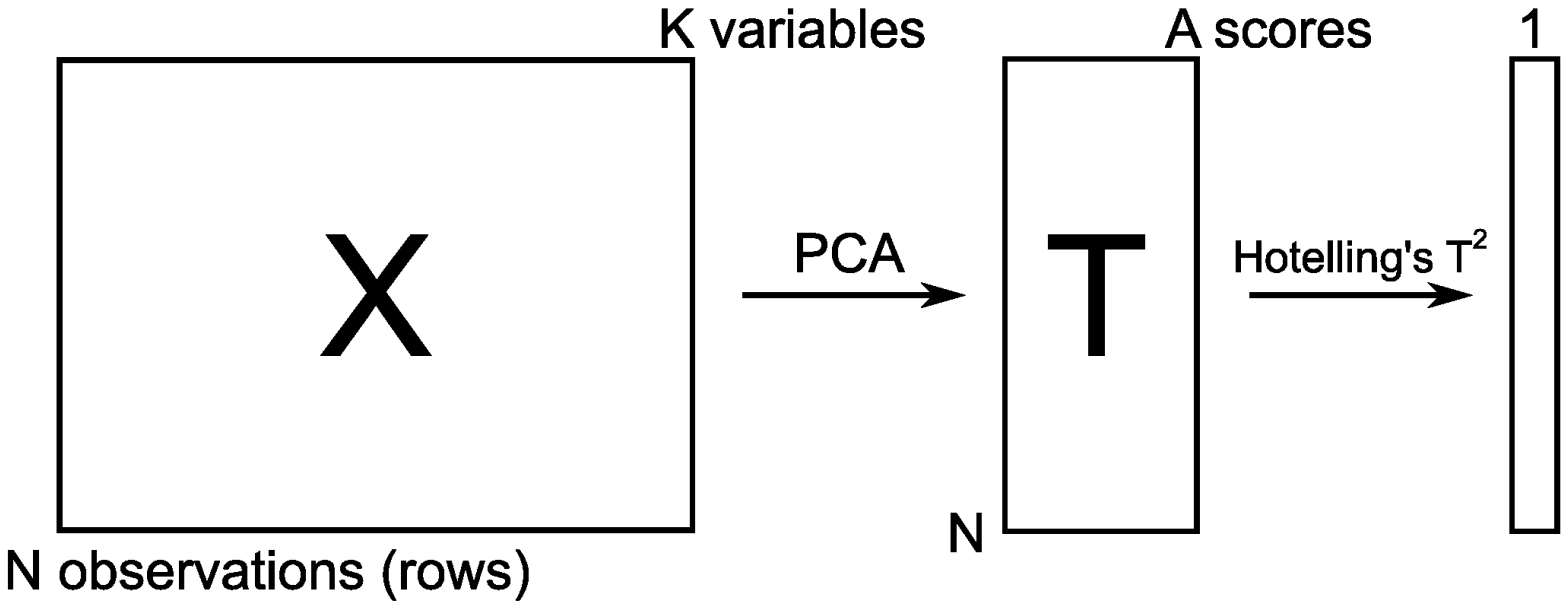
Why we use Hotelling's T^2

Resume from last class: *slides 28 to 32*

Unfortunately, I've added some more details, and rearranged the slides

Hotelling's T^2

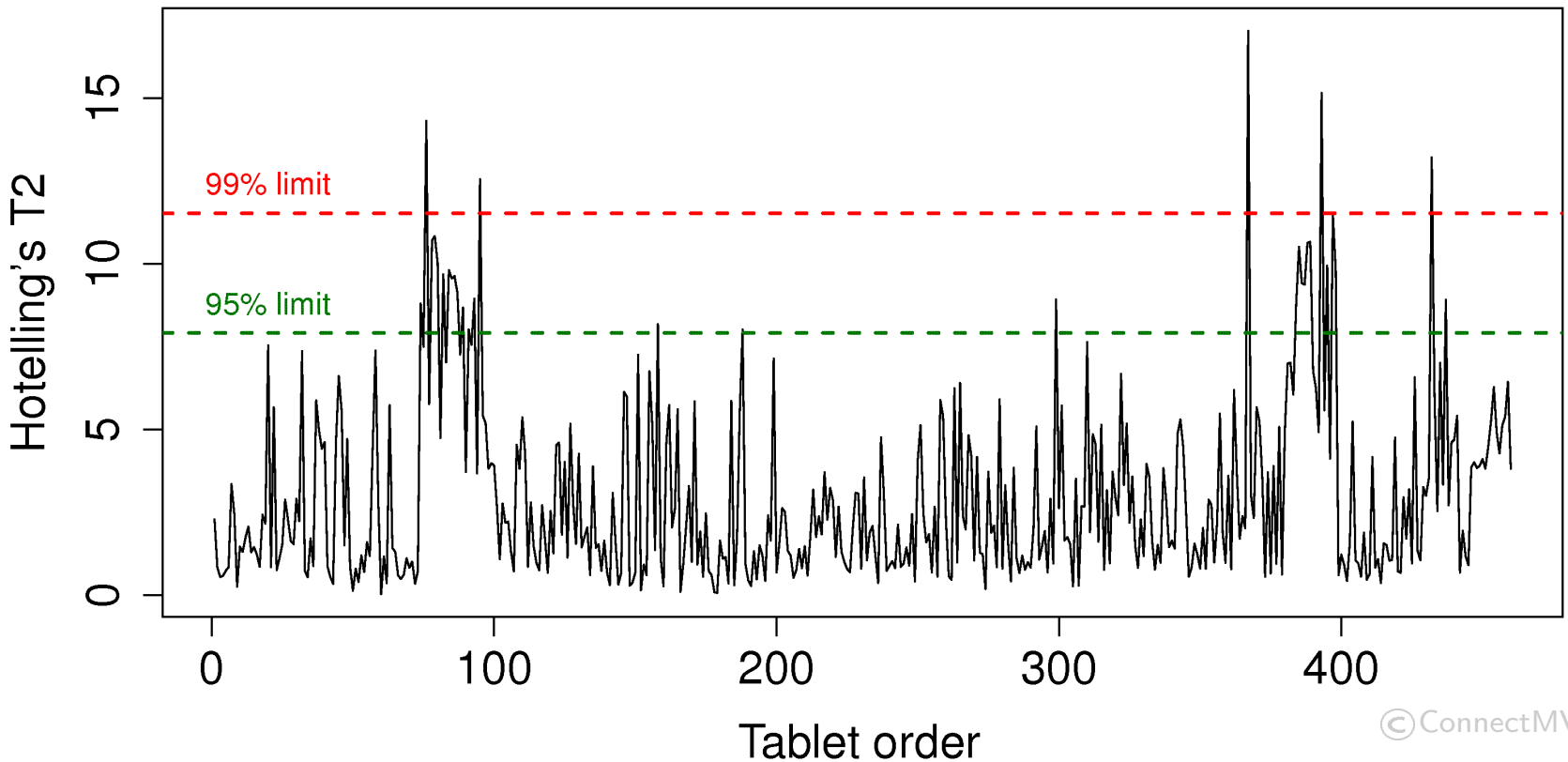
- After extracting components from **X** we accumulate A score vectors in matrix **T**



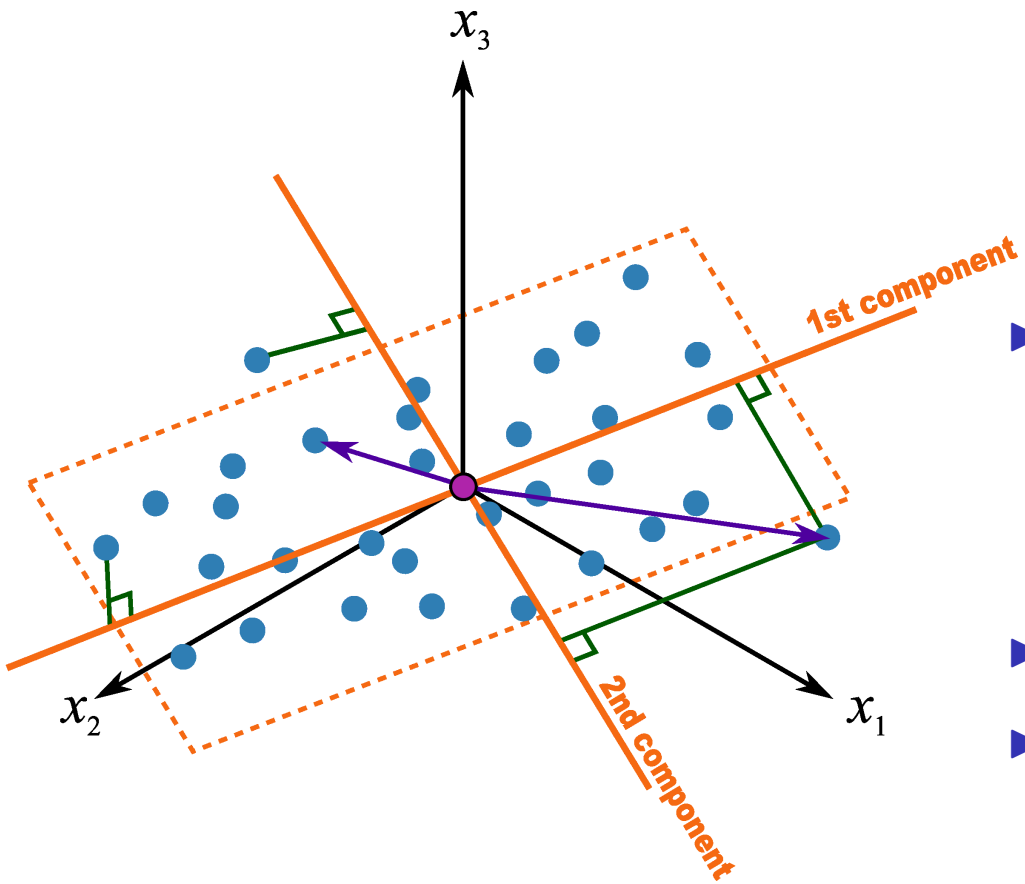
- T_i^2 is a summary of all A components within row i
- $T_i^2 = \sum_{a=1}^{a=A} \left(\frac{t_{i,a}}{s_a} \right)^2$
- s_a = standard deviation of score column a

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

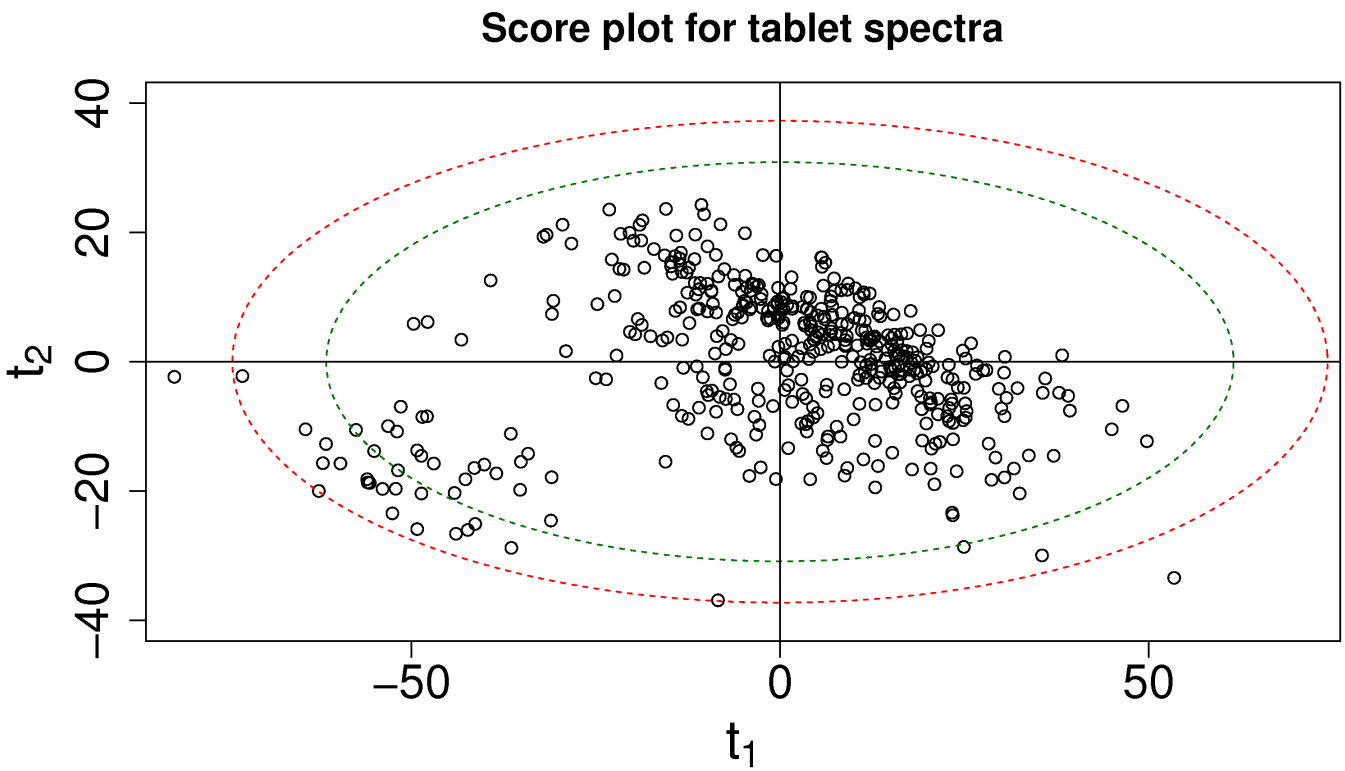


$$T_i^2 = \sum_{a=1}^{a=A} \left(\frac{t_{i,a}}{s_a} \right)^2 \geq 0$$

- 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 $T^2_{A,\alpha=0.05}$

Hotelling's T^2

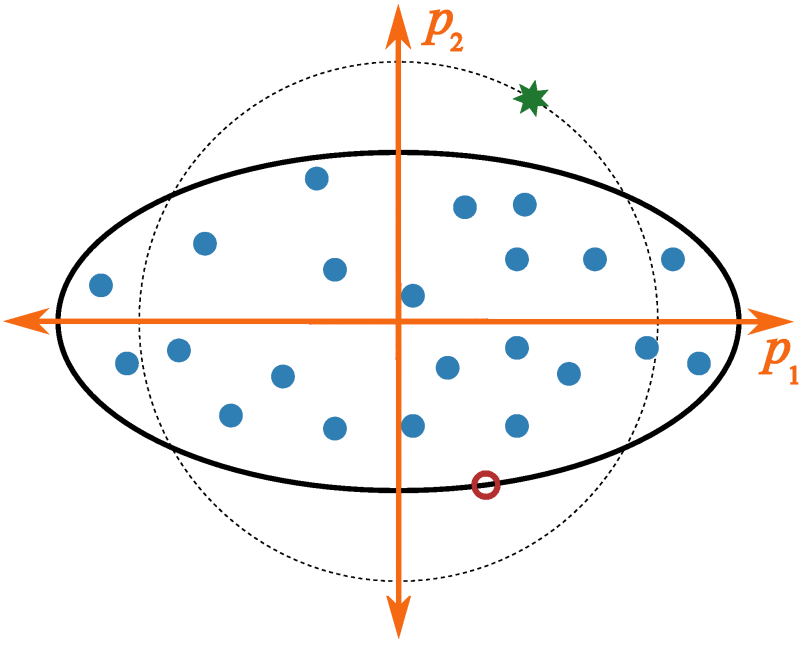
- If $A = 2$, equation for 95% limit $= T^2_{A=2,\alpha=0.05} = \frac{t_1^2}{s_1^2} + \frac{t_2^2}{s_2^2}$
- An equation for an ellipse
- s_1 and s_2 are constant for a given model
- Points on ellipse have a constant distance from model center

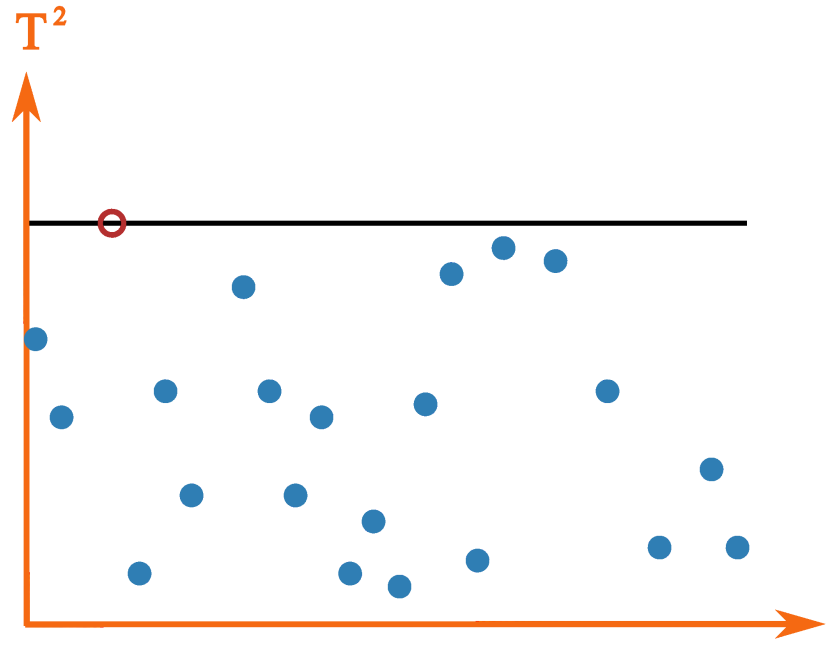


Hotelling's T^2

- Hotelling's T^2 = distance of every point from center, taking (co)variance into account
- Why not use a Euclidean distance $T_i^2 = \sum_{a=1}^{a=A} \left(\frac{t_{i,a}}{1} \right)^2$
- Instead we use the Mahalanobis distance:
  $T_i^2 = \sum_{a=1}^{a=A} \left(\frac{t_{i,a}}{s_a} \right)^2 \geq 0$

Why Euclidean distances don't work





The green point is equidistant from the center, but doesn't accurately reflect "outlyingness"

The same red point however is "equally far" from the model center, at all points on the ellipse

Inspiration for left image is due to Rasmus Bro's video:

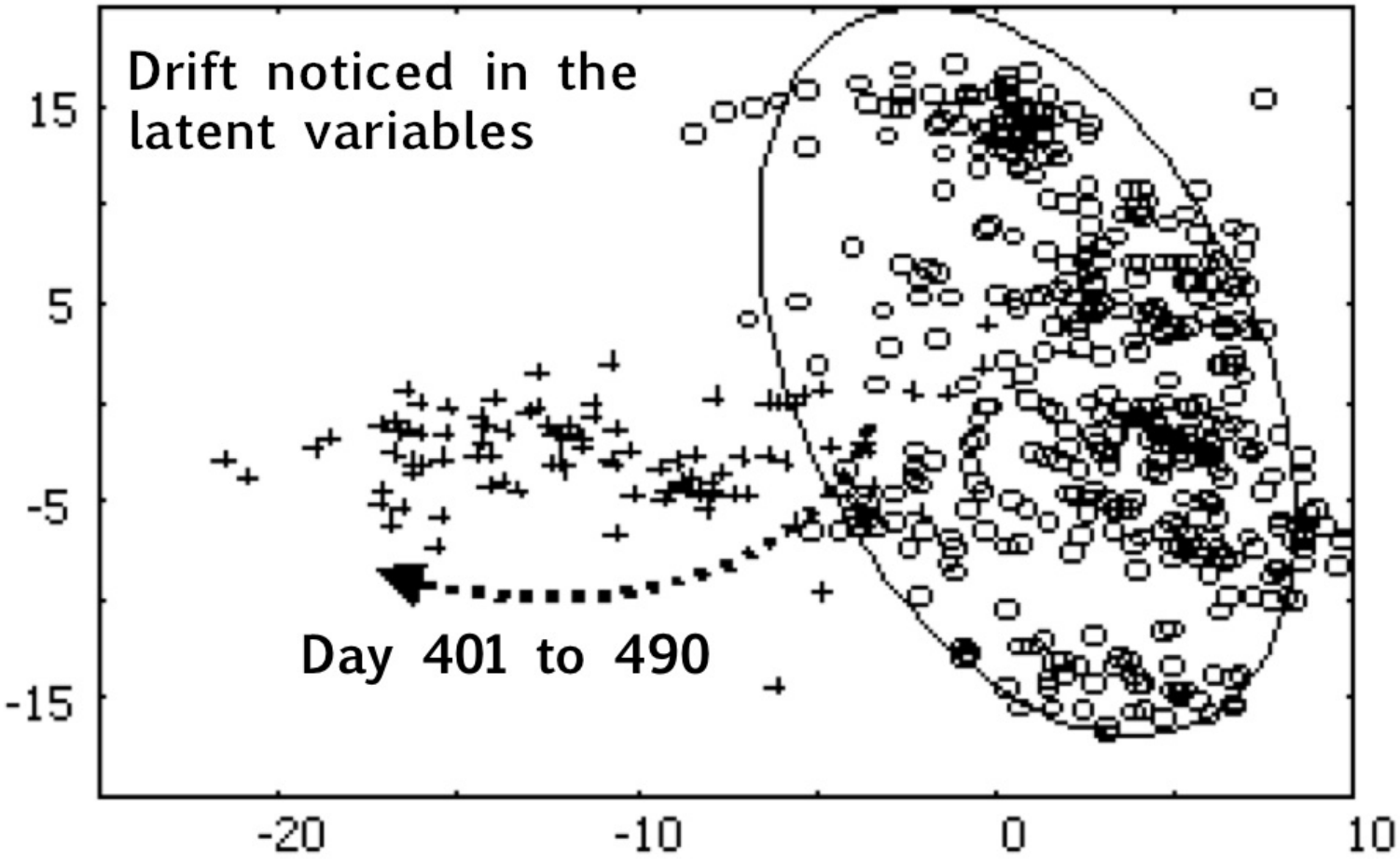
http://www.youtube.com/watch?v=ExoAbXPJ7NQ

Contribution plots

Resume from last class: *slides 60 to 66*

Unfortunately, I've added some more details, and rearranged the slides

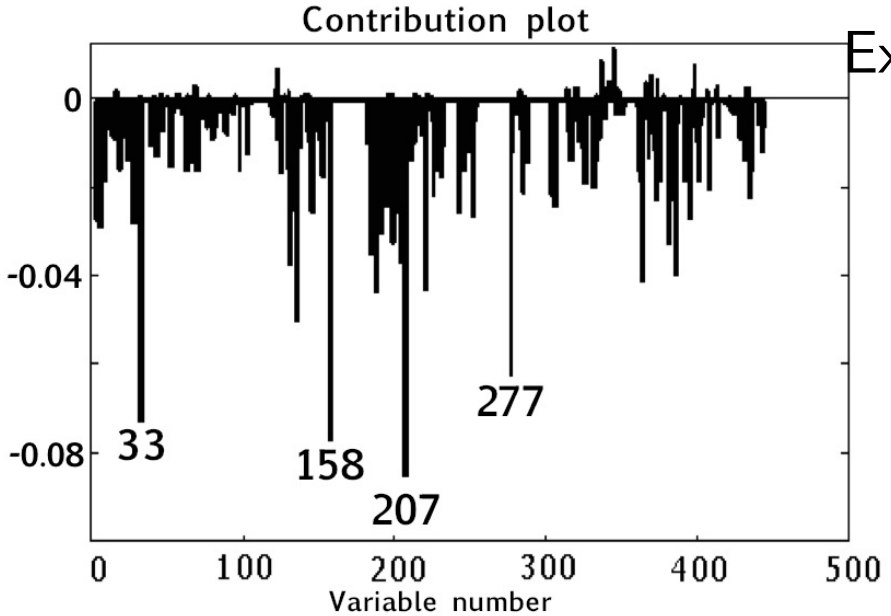
Diagnosing a problem



- Interrogate the latent variables to see what changed

LVM for troubleshooting: contribution plot

- Shows difference between two points in the score plot

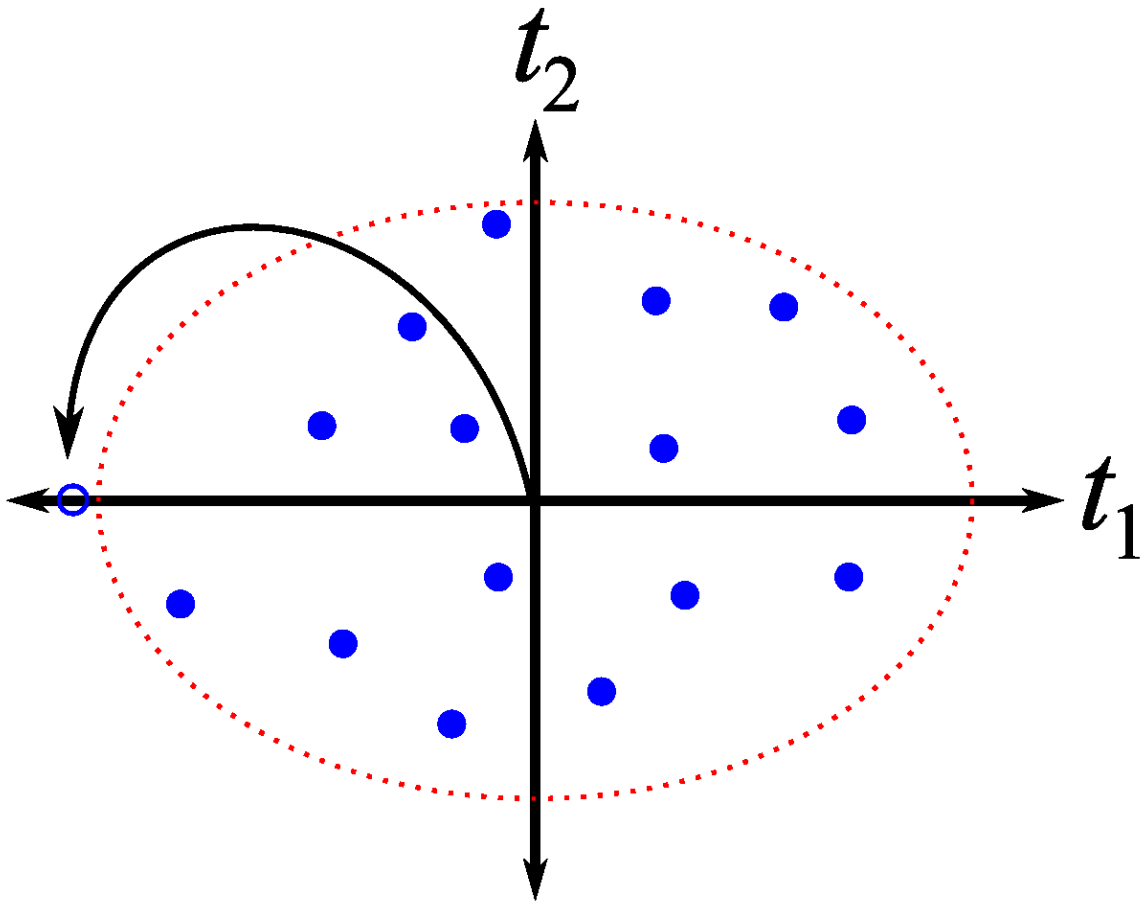


Example:

- **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

Contributions in the score space: one PC



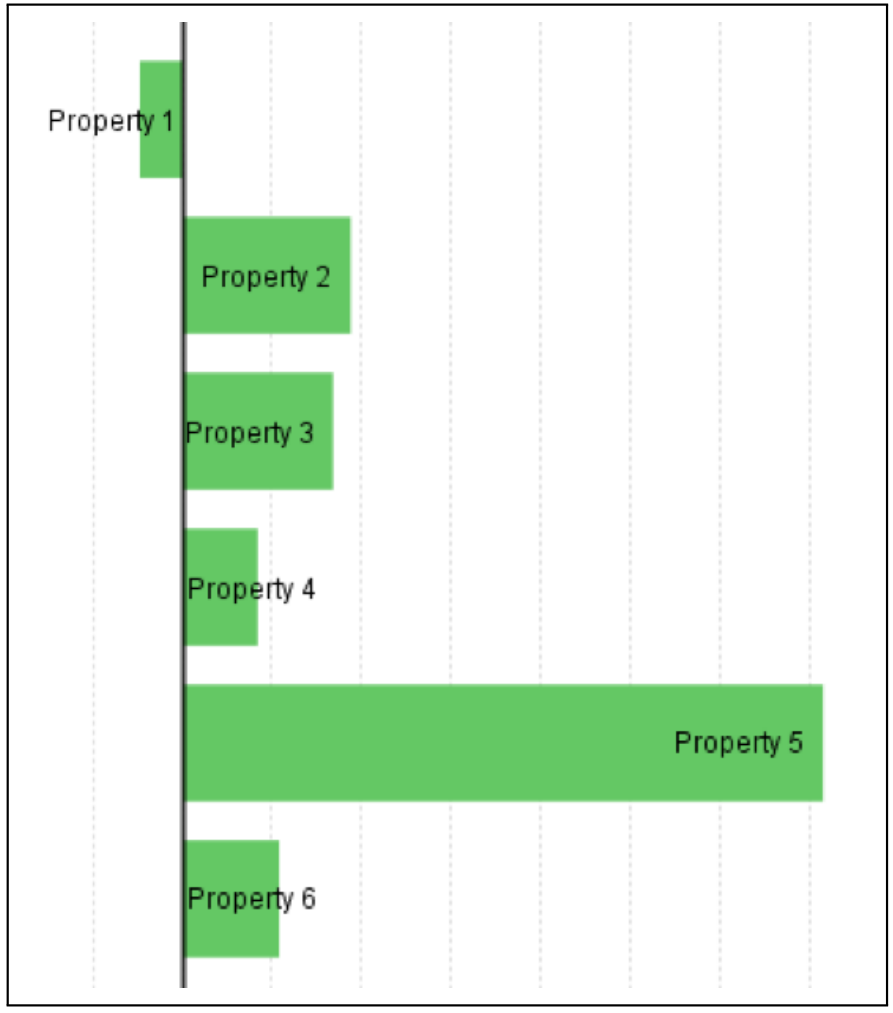
From the model center to a point

Contributions in the score space: one PC

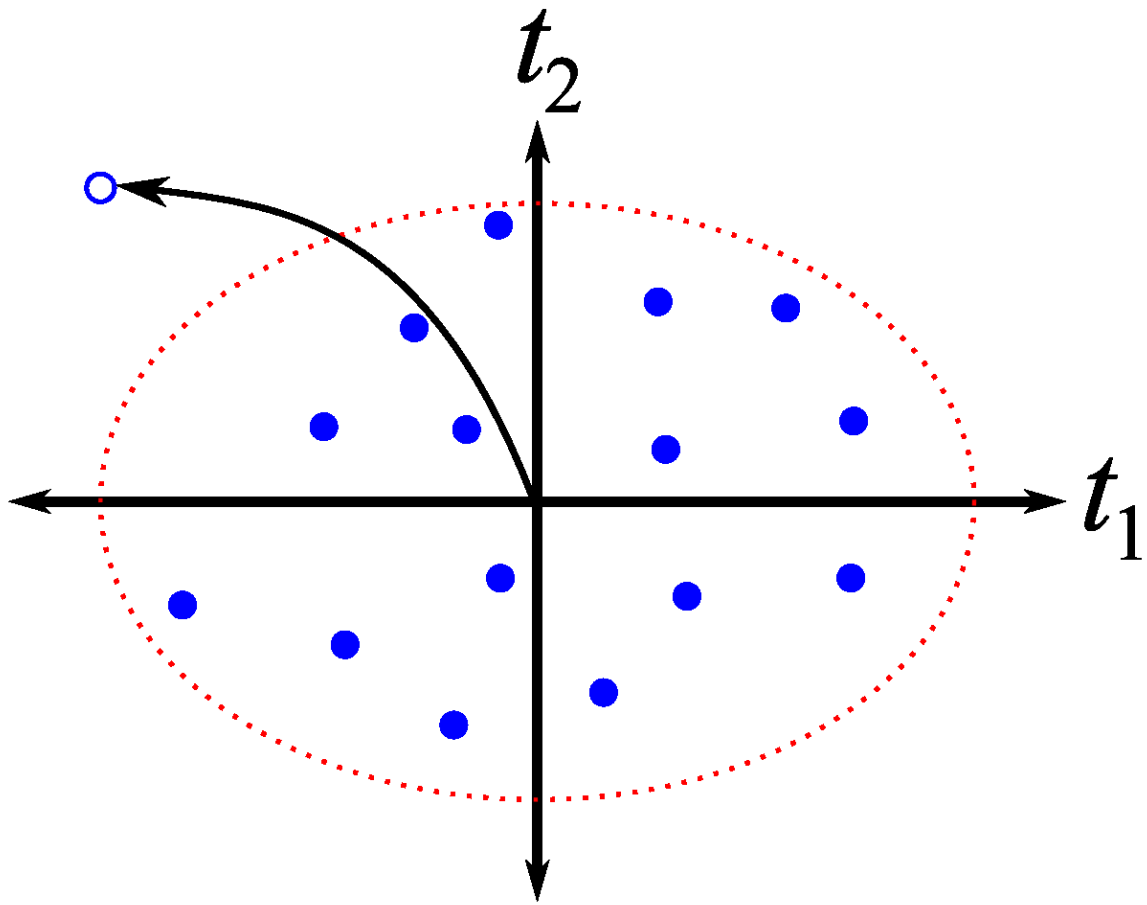
Score $= t_{i,a} = \mathbf{x}_i \mathbf{p}_a =$ linear combination

- $\begin{bmatrix} x_{i,1}p_{1,a} & x_{i,2}p_{2,a} & \dots & x_{i,K}p_{K,a} \end{bmatrix}$ $\longleftarrow$ there are K terms

- relative size of terms is interpreted
- most often shown as a bar plot
- absolute value on y-axis is *never* used/not shown
- not sensible to interpret contributions for observation with a small score
- example here has $K = 6$
- signs can be interpreted, but rather verify in raw data



Contributions in more than 1 score



From the model center to a point

Contributions in more than 1 score

Summation of the contributions from each score, weighted by the size of the score.

Consider PC1 and PC2 for variable k:

- contribution in t_1 direction $= x_{i,k} p_{k,1}$
- contribution in t_2 direction $= x_{i,k} p_{k,2}$
- joint contribution $= x_{i,k} \left| p_{k,1} \cdot \frac{t_{i,1}}{s_1} \right| + x_{i,k} \left| p_{k,2} \cdot \frac{t_{i,2}}{s_2} \right|$

In general: joint contribution for variable $x_k =$

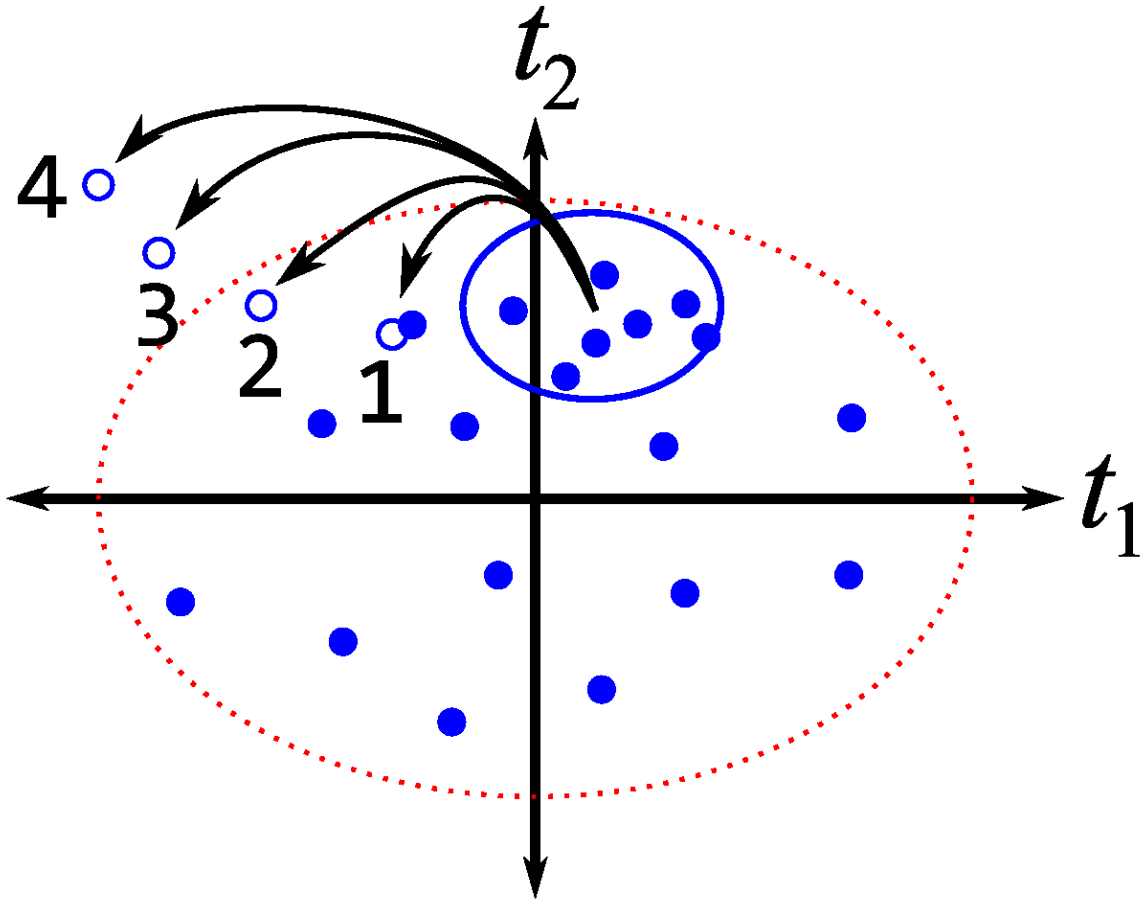
$$\text{contrib}(x_k) = x_{i,k} \sqrt{\sum_a \left(p_{k,a} \cdot \frac{t_{i,a}}{s_a} \right)^2}$$

Contribution plots in T^2

Not uniform in various software:

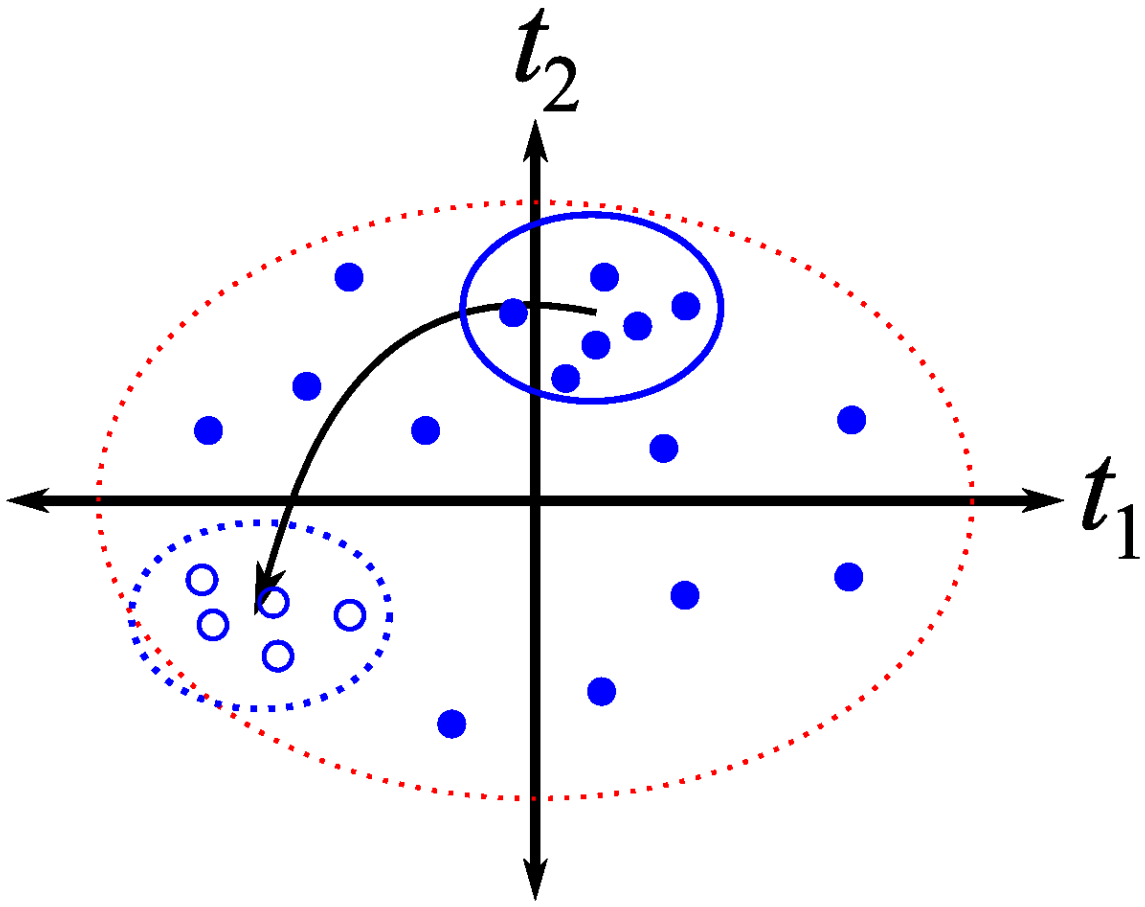
- Cleanest: use the weighted sum of score contributions, as shown before.
- Alvarez *et al.* - paper 21
- Kourti and MacGregor - paper 81
- Mason, Tracy and Young: "Decomposition of T^2 for multivariate control chart interpretation", *Journal of Quality Technology*, **27**, 99-108, 1995.

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

Contributions: modifying the starting point

We can modify the starting point, not necessary to use origin:

- $t_{i,a}^{(\text{to})} = \mathbf{x}_i^{(\text{to})} \mathbf{p}_a$
- $t_{i,a}^{(\text{from})} = \mathbf{x}_i^{(\text{from})} \mathbf{p}_a \qquad \longleftarrow \qquad$ usually the origin: $t_{i,a}^{(\text{from})} = 0$

Subtract:

$$t_{i,a}^{(\text{to})} - t_{i,a}^{(\text{from})} = \left(\mathbf{x}_i^{(\text{to})} - \mathbf{x}_i^{(\text{from})} \right) \mathbf{p}_a$$

$$\Delta t_{i,a} = \Delta \mathbf{x}_i \mathbf{p}_a \qquad \longleftarrow \text{ plot as bar plot}$$

In general:

$$\text{contrib}(x_k) = \left(x_{i,k}^{(\text{to})} - x_{i,k}^{(\text{from})} \right) \sqrt{\sum_a \left(p_{k,a} \cdot \frac{t_{i,a}^{(\text{to})} - t_{i,a}^{(\text{from})}}{s_a} \right)^2}$$

Contributions in the residuals

SPE $= \mathbf{e}_i' \mathbf{e}_i$ where $\mathbf{e}_i' = \mathbf{x}_i' - \widehat{\mathbf{x}}_i'$

- $\left[(x_{i,1} - \hat{x}_{i,1}) \quad (x_{i,2} - \hat{x}_{i,2}) \quad \ldots \quad (x_{i,K} - \hat{x}_{i,K})\right]$ $\longleftarrow$ bar plot
- Could show squared values: $(x_{i,k} - \hat{x}_{i,k})^2$ for variable k
- But sometimes $+$ve and $-$ve patterns in the bars are helpful to identify the fault signature
- See work of Yoon and MacGregor on fault signatures
- Don't interpret absolute value of the error bars
- Don't interpret contributions for observations with small SPE
- Large bar: doesn't always mean that variable is a problem (*example on board*)

Contribution plots: T^2 and SPE

Joint T^2 and SPE monitoring plots

- *Illustrated on the board*

Leverage

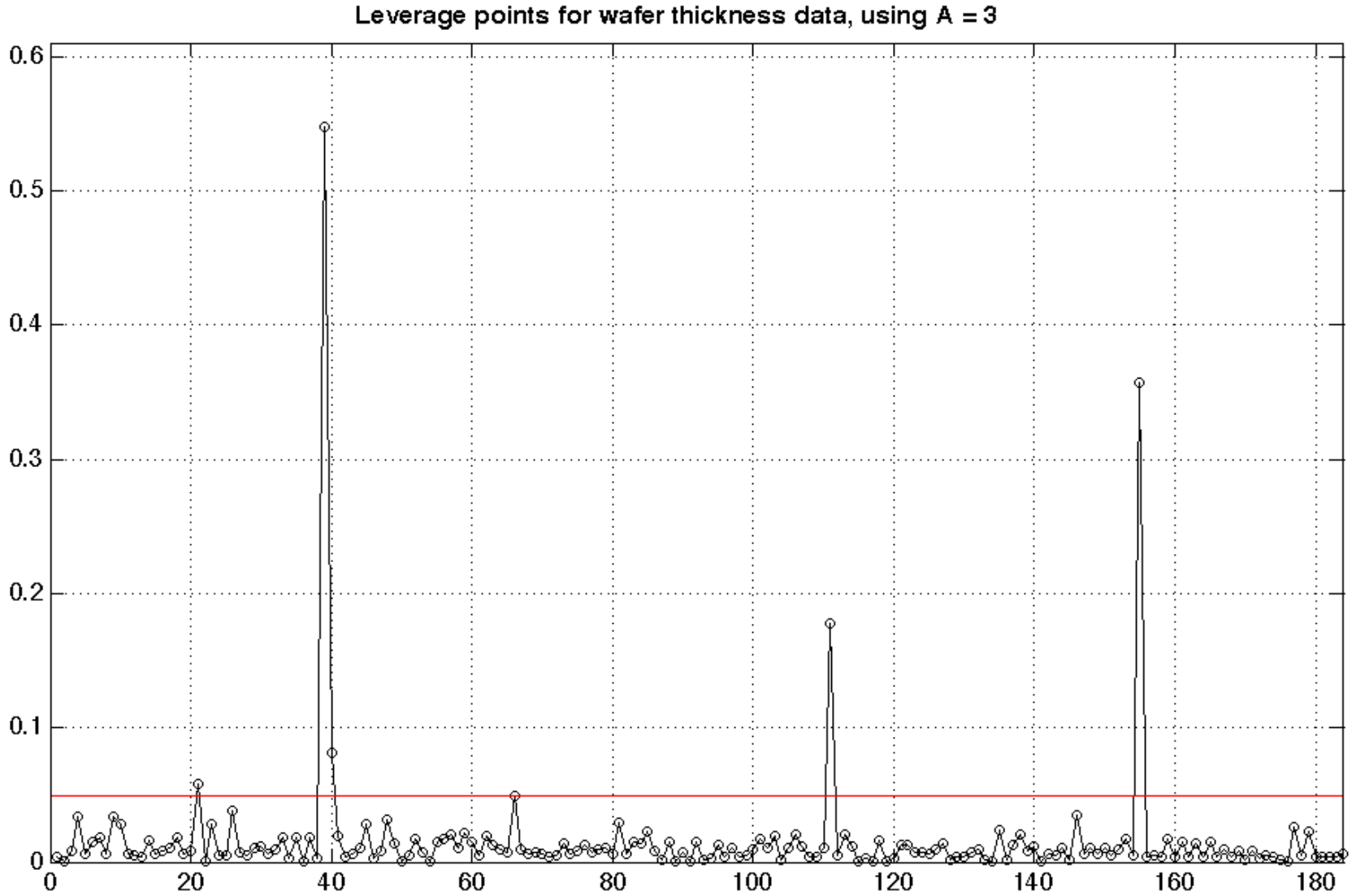
You might see the concept of "leverage" in software packages:

Each observation has leverage on the mode

$$\text{Leverage}_i = \text{diag}\left\{\mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\right\}_{(i,i)} > 0$$

- $(\mathbf{T}'\mathbf{T}) =$
- $\text{Leverage}_i =$ scaled down version of T_i^2
- $\sum_{i=1}^{i=N} \text{Leverage}_i = A =$ the number of columns in $\mathbf{T}$
- Cut off for $\text{Leverage}_i = 3 \cdot \frac{A}{N}$
- Points with $\text{Leverage}_i >$ cut off have large influence on model

Leverage example



$$\text{Cut off} = 3 \cdot \frac{A}{N} = 3 \times 3/184$$

Variable importance to prediction

Characteristics of variables that have important role in model?

- Have large (absolute) weights: why?
- Come from a component that has a high R^2

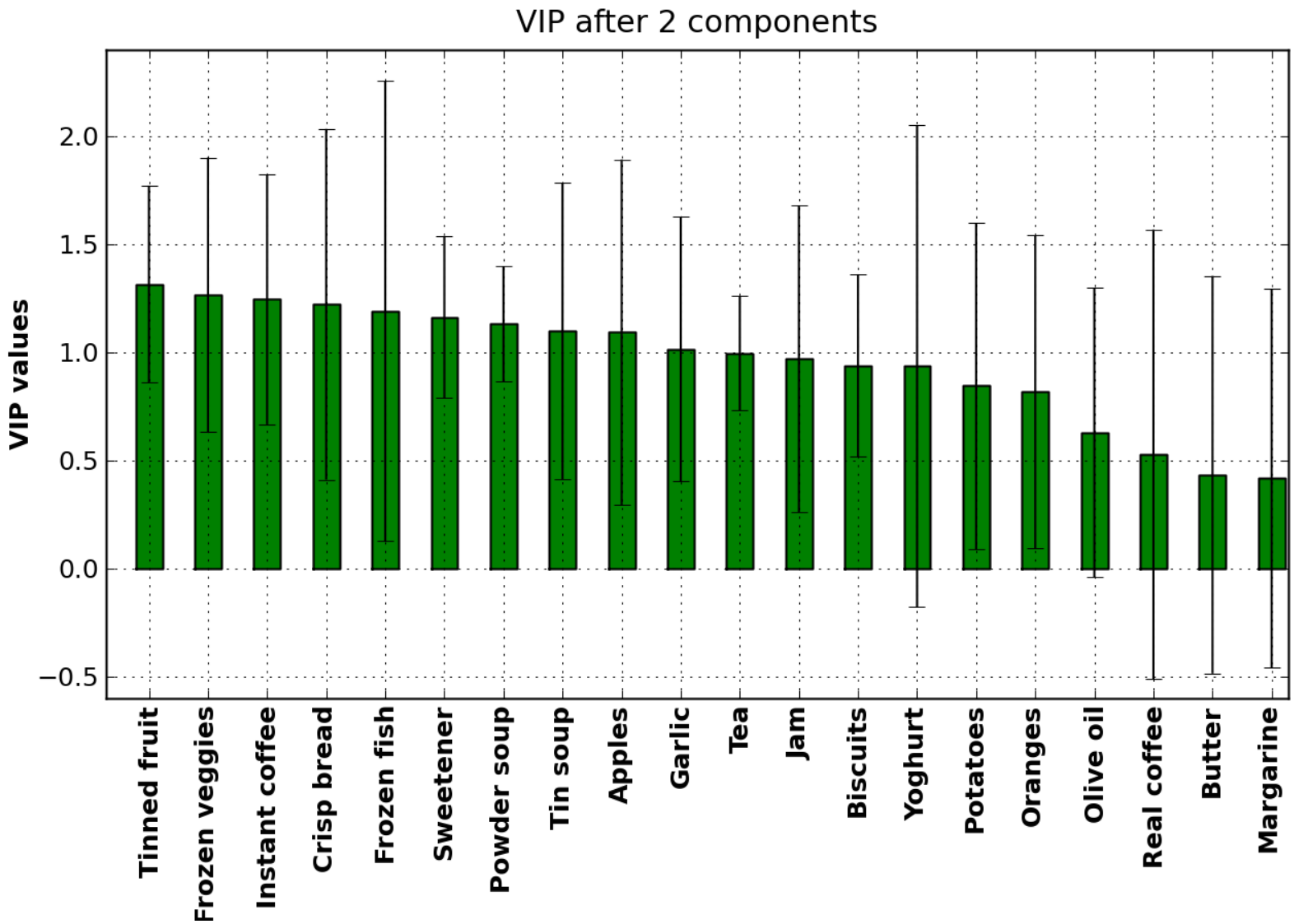
Combining these two concepts we calculate *for each variable*:

Importance of variable k using A components

$$VIP^2_{A,k} = \frac{K}{\mathrm{SSX}_0 - \mathrm{SSX}_A} \cdot \sum_{a=1}^{A} (\mathrm{SSX}_{a-1} - \mathrm{SSX}_a)\, P^2_{a,k}$$

- SSX_a = sum of squares in the $\mathbf{X}$ matrix after a components
- $\frac{\mathrm{SSX}_{a-1} - \mathrm{SSX}_a}{SSX_A}$ = incremental R^2 for a^{th} component
- $\frac{\mathrm{SSX}_0 - \mathrm{SSX}_A}{SSX_A} = R^2$ for model using A components
- Messy, but you can show that $\sum_k VIP^2_{A,k} = K$
- Reasonable cut-off =

Variable importance to prediction



Jackknifing

We re-calculate the model $G + 1$ times during cross-validation:

- G times, once per group
- The "+1" is from the final round, where we use **all** observations

We get $G + 1$ estimates of the PCA model parameters:

- loadings
- VIP values

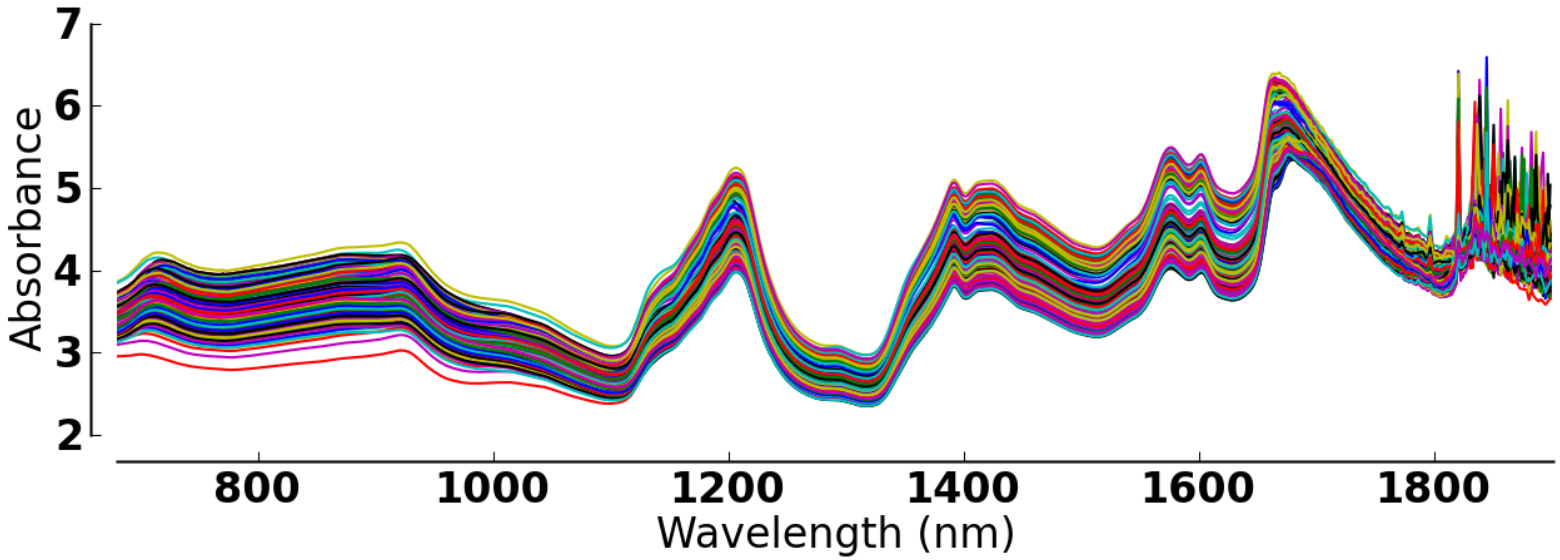
for every variable $(1, 2, \ldots K)$.

Can now calculate confidence intervals (caution with CI on loadings)

- Martens and Martens (paper 43) describing jackknifing.
- Efron and Tibshirani describe the bootstrap and jackknife.

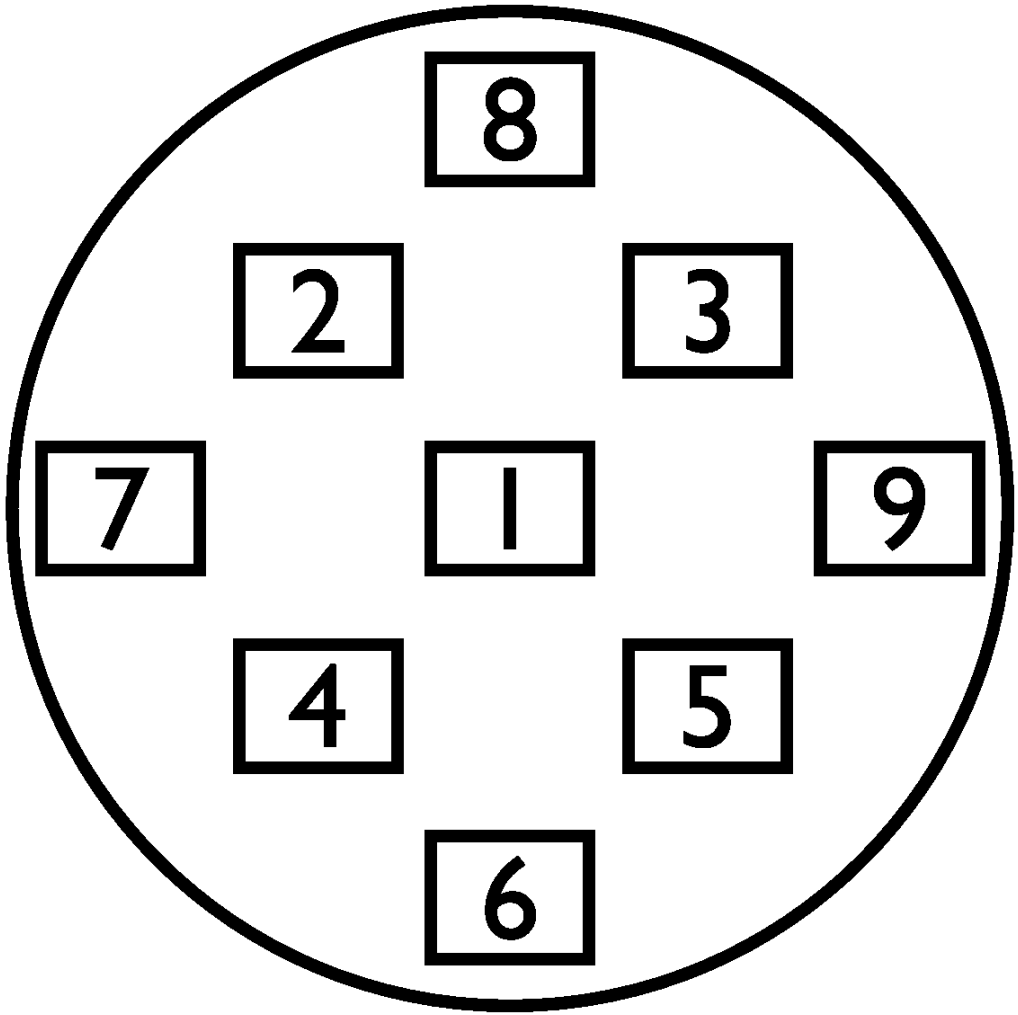
Case studies

- Raw material characterization
- Near infra-red spectra of tablets



Wafer case study

- Data source: Silicon wafer thickness
- Nine thickness measurements from a silicon wafer.
- Thickness measured at the nine locations



Wafer case study I

1. Build a PCA model on the data on the first 100 rows.
2. Plot the scores. What do you notice?
3. Investigate the outliers with the contribution tool.
4. Verify that the outliers exist in the raw data
5. Exclude any unusual observations and refit the model
6. Did you get all the outliers? Check the scores and SPE. Repeat to get all outliers removed.
7. Plot a loadings plot for the first component. What is your interpretation of p_1?
8. Given the R^2 and Q^2 values for the first component, what is your interpretation about the variability in this process? (Remember the goal of PCA is to explain variability)

Wafer case study II

9. What is the interpretation of p_2? From a quality control perspective, if you could remove the variability due to p_2, how much of the variability would you be removing from the process?
10. Plot the corresponding time series plot for t_1. What do you notice in the sequence of score values?
11. Repeat the above question for the second component.
12. Use all the data as testing data (184 observations, of which the first ≈ 100 were used to build the model).
13. Do the outliers that you excluded earlier show up as outliers still? Do the contribution plots for these outliers give the same diagnosis that you got before?
14. Are there any new outliers in points 101 to 184? If so, what are is their diagnosis?

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 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

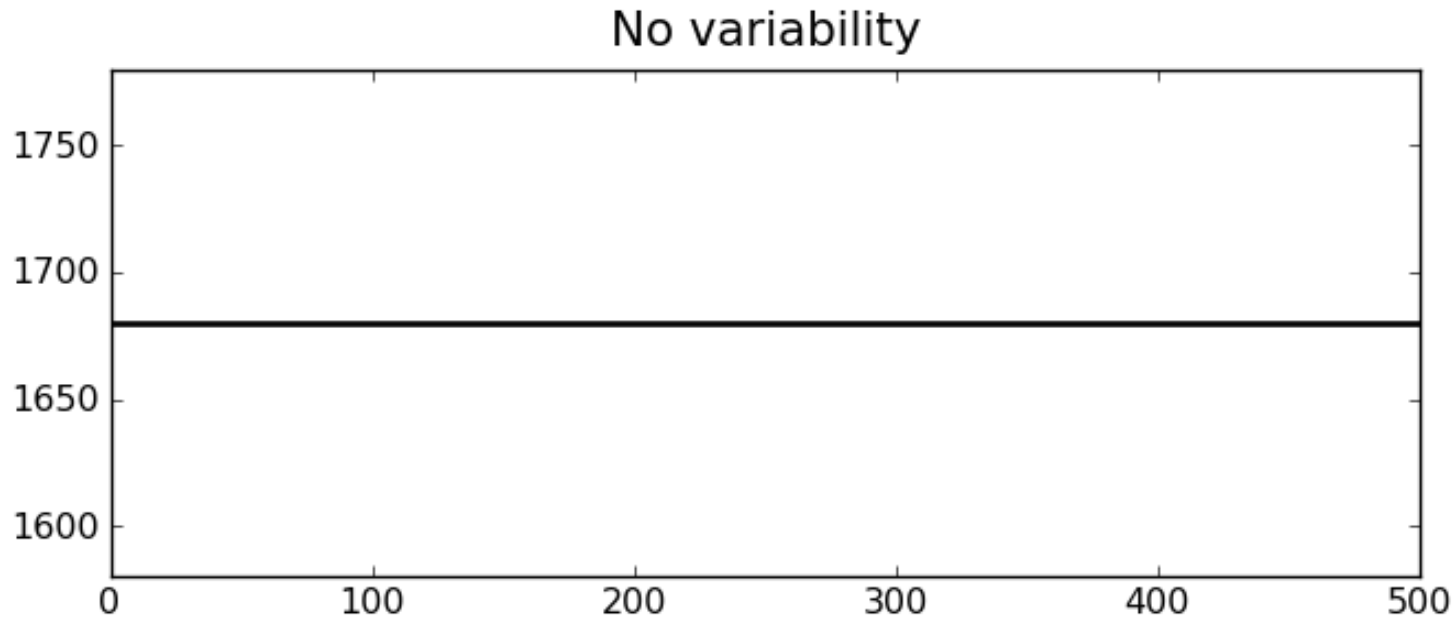
Monitoring analogy: making errors

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

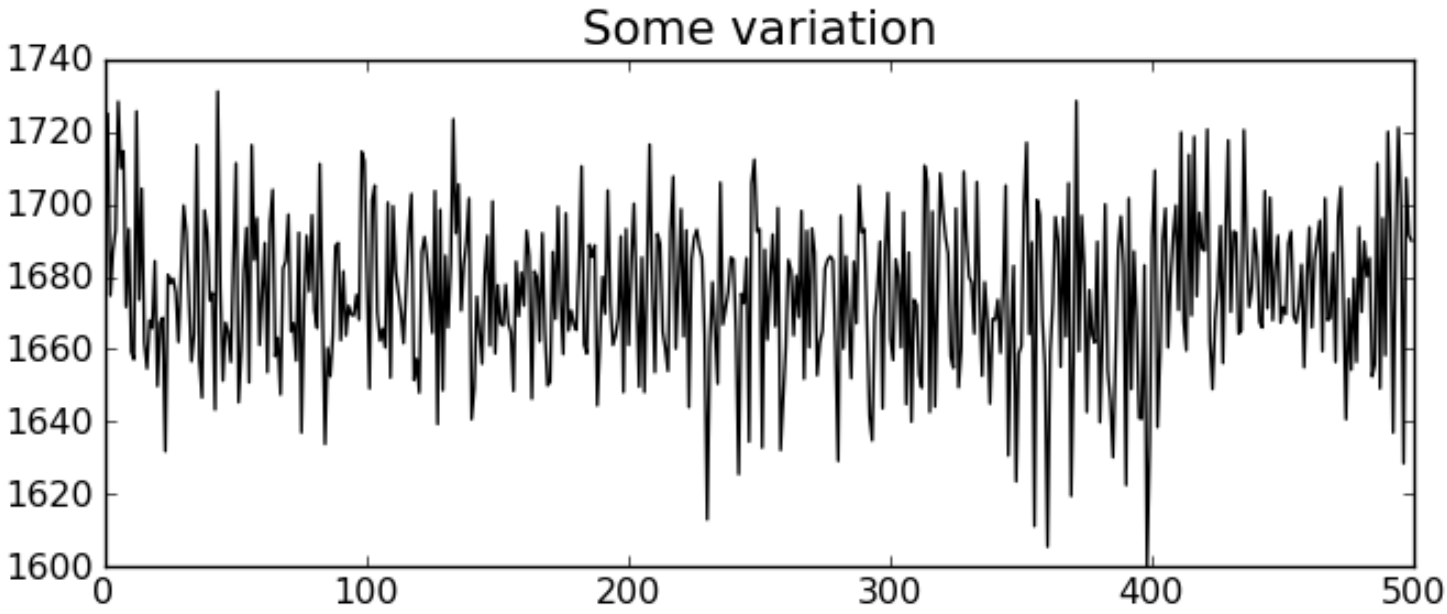
- **Type 2 error**: *you feel OK*; but go to doctors for physical and they detect a problem
  - You feel within your limits,
  - but the truth is: "*you are not healthy*"
  - **Type 2 error** = don't raise an alarm when there is a problem
- The grid

Monitoring concept for a process

Our goal: We want process stability



Variability

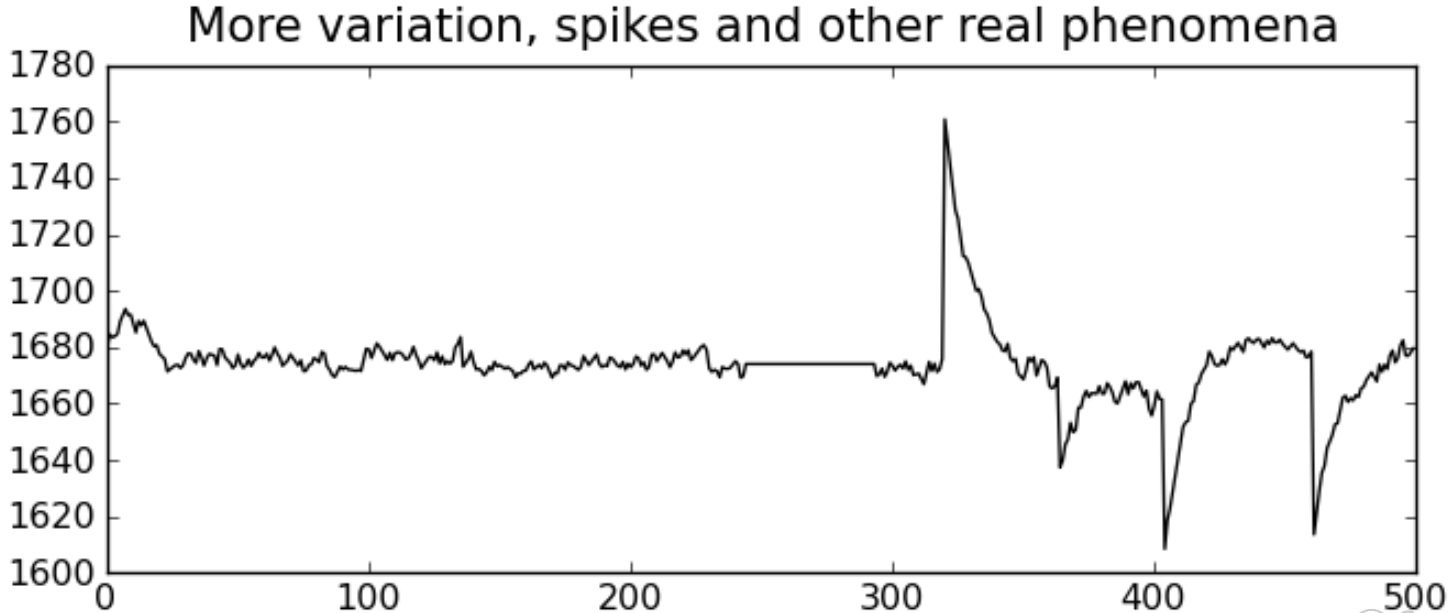


Best case: we have **unaccounted sources** of noise: called **error**

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



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

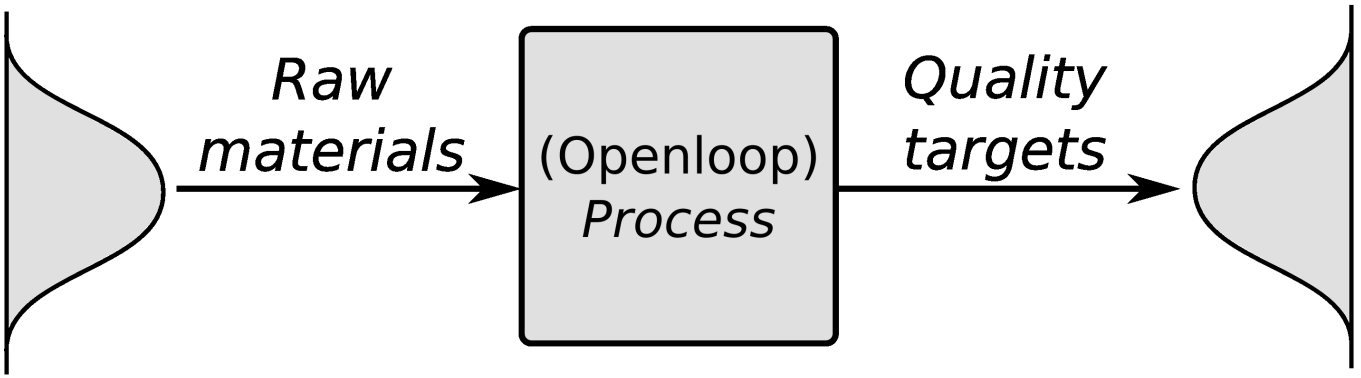
Variability costs you money

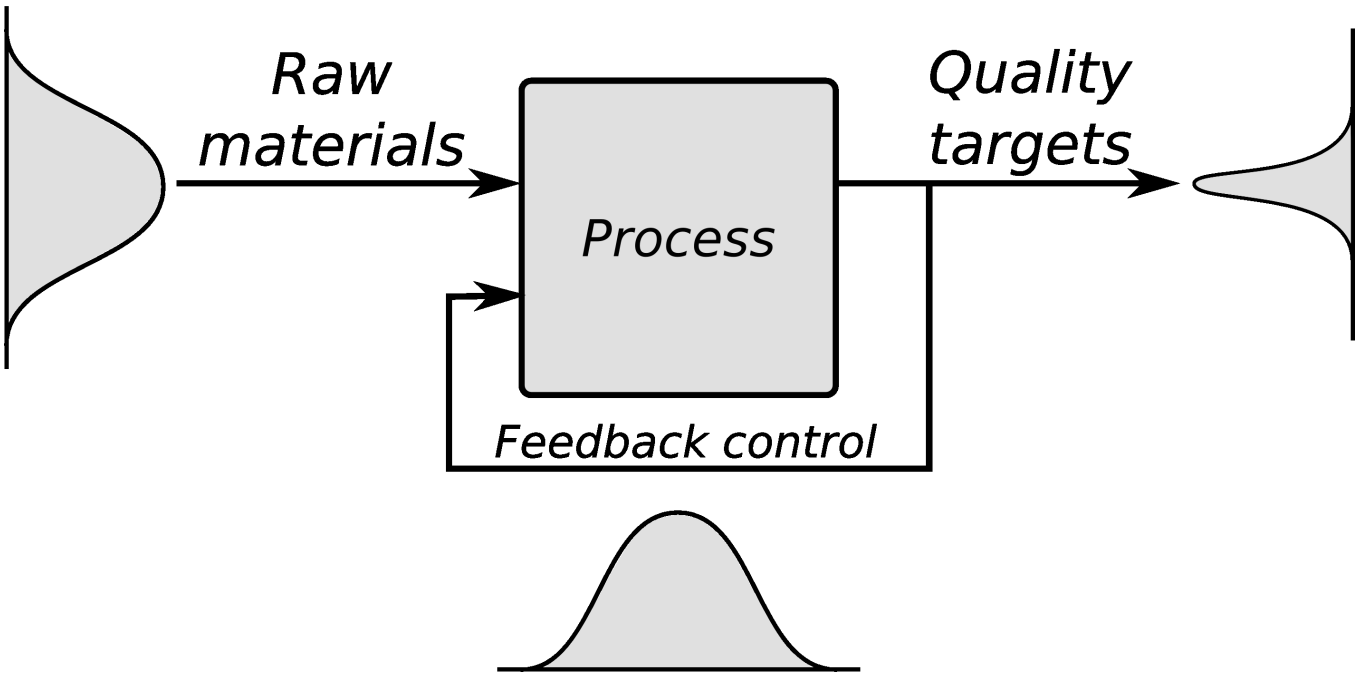
The high cost of variability in your final product:

1. Inspection costs:
   - high variability: test every product (expensive, inefficient, sometimes destructive)
   - low variability: limited inspection required
2. Off-specification products cost you, and customer, money:
   - reworked
   - disposed
   - sold at a loss

The high cost of variability in your raw materials

- Flip it around: you receive highly variable raw materials:
  - That variability lands up in your product, or
  - you incur additional cost (energy/time/materials) to process it





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: relationship to feedback control

- "Process monitoring" also called "Statistical Process Control" (SPC)
- We will avoid this term due to potential confusion:
- Monitoring is *similar* to (feedback) control:
  - continually applied
  - checks for deviations (error)
- Monitoring is *different* to (feedback) control:
  - adjustments are **infrequent**
  - usually **manual**
  - adjust due to **special causes**

- Process monitoring: make *permanent* adjustments to reduce variability
- Feedback control: *temporarily* compensates for the problem

Other types of monitoring you will see

Monitoring is widely used in all industries

- Managers: monitor geographic regions for hourly sales, downtime, throughput
- Engineers: monitor large plants, subsections, and unit operations

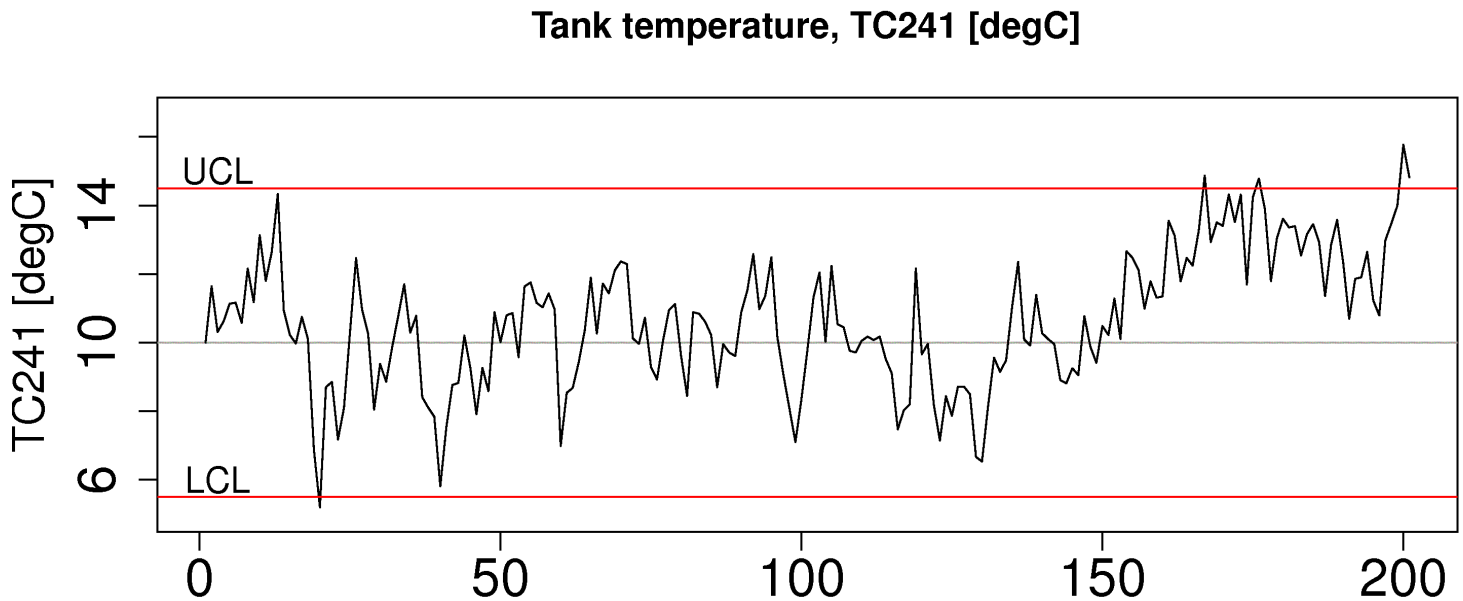
Tools/buzzwords used go by names such as:

- Dashboards
- Analytics
- BI: business intelligence,
- KPI: key performance indicators

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

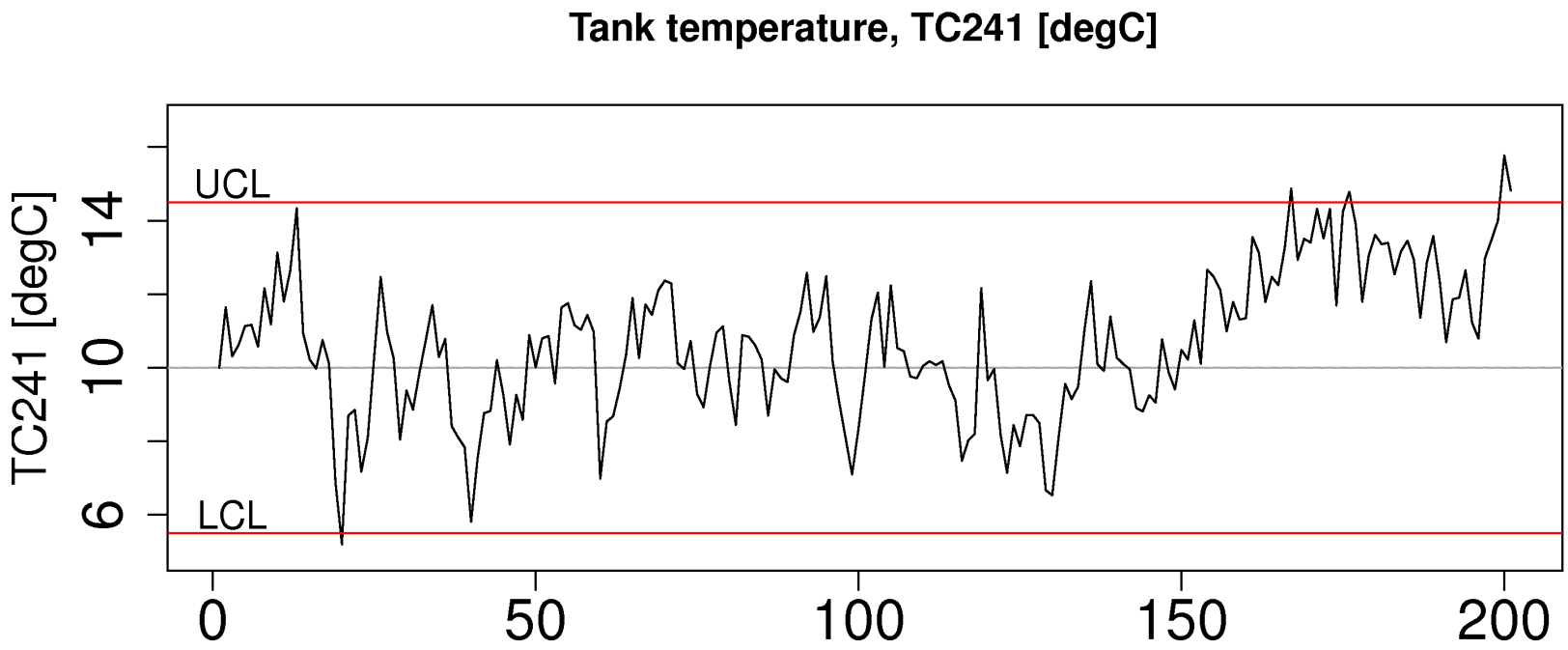
Key point

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

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

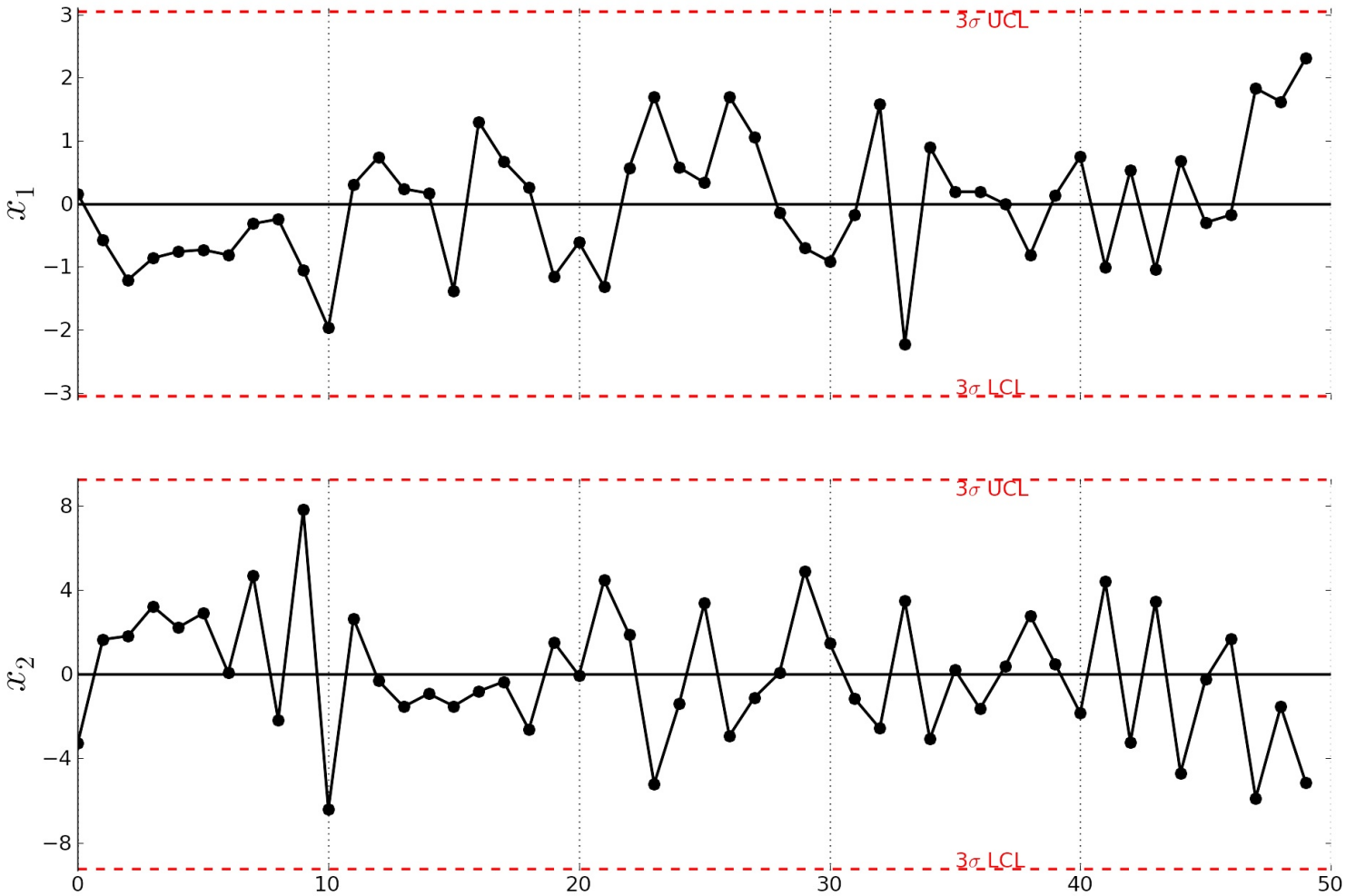
- **But, you cannot simultaneously have low type I and type II error**

Discussion



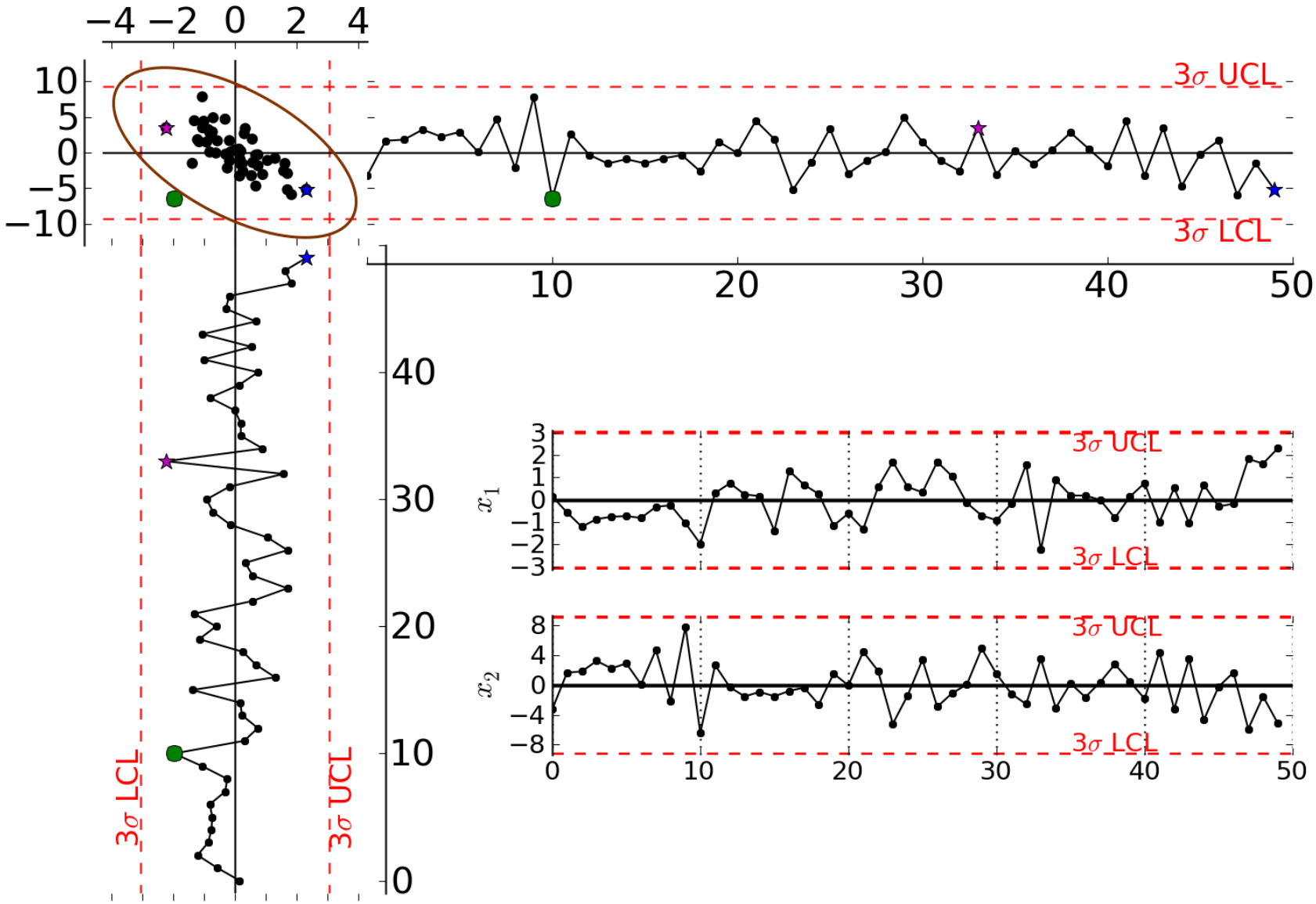
1. What action is taken when outside the limits
2. What if data goes missing?

Discussion



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

Discussion: multivariate monitoring



Discussion: monitoring only final quality data

Lab measurements have a long time delay:

- process already shifted by the time lab values detect a problem (continuous)
- batches have to placed on hold until lab results return
- very hard to find cause-and-effect for diagnosis
  - e.g. low product strength could be caused by multiple reasons

Discussion: monitoring only final quality data

Measurements from real-time systems are:

- available more frequently (less delay) than lab measurements
- often are more precise, often with lower error
- more meaningful to the operating staff
- contains almost unique "fingerprint" of problem (helps diagnosis)
  - Now we can figure out what caused low product strength

"Variables" monitored don't need to be from on-line sensors: could be a calculated value

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

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

Process monitoring with PCA: Hotelling's T^2

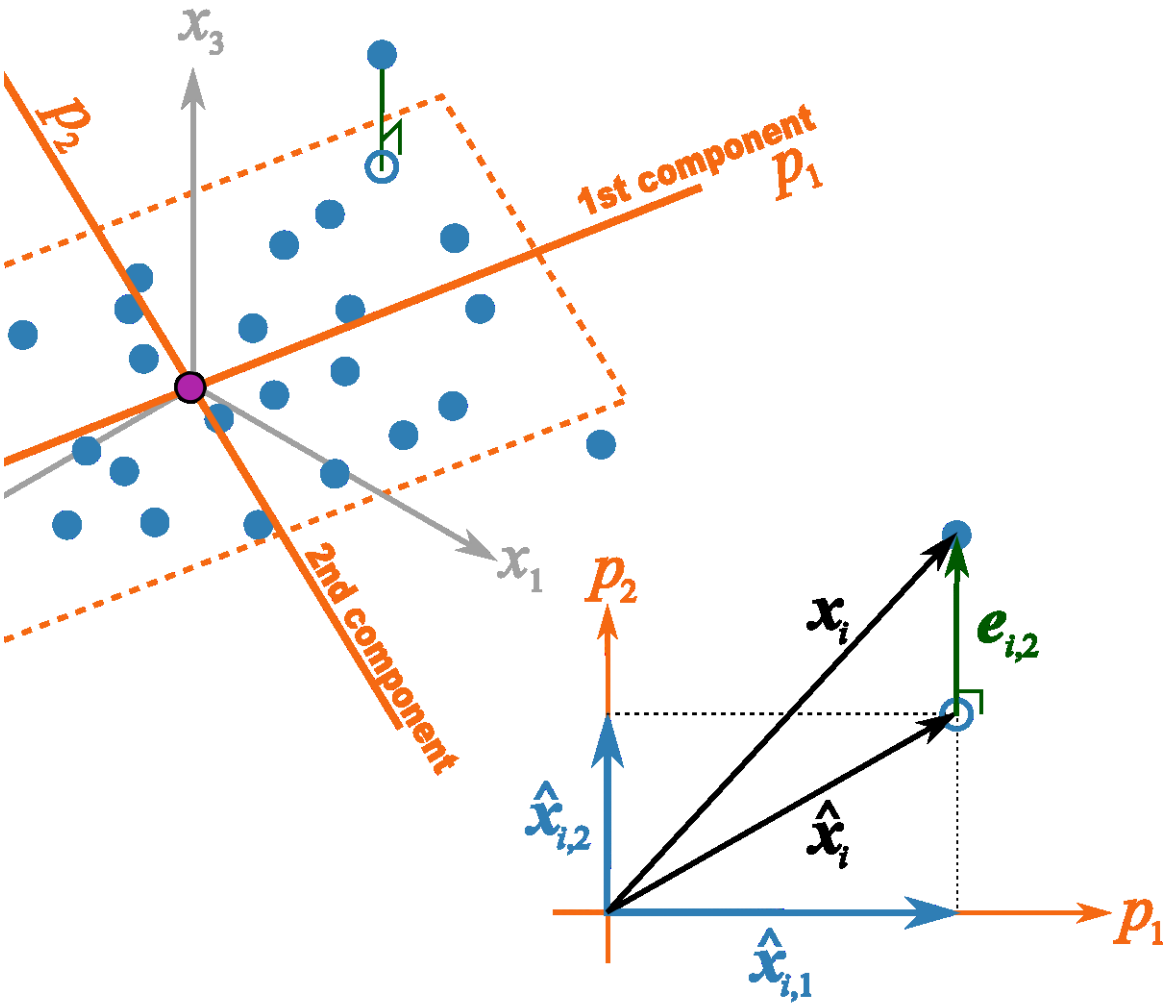
$$\text{Hotelling's } T^2 = \sum_{a=1}^{a=A} \left(\frac{t_a}{s_a} \right)^2$$

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

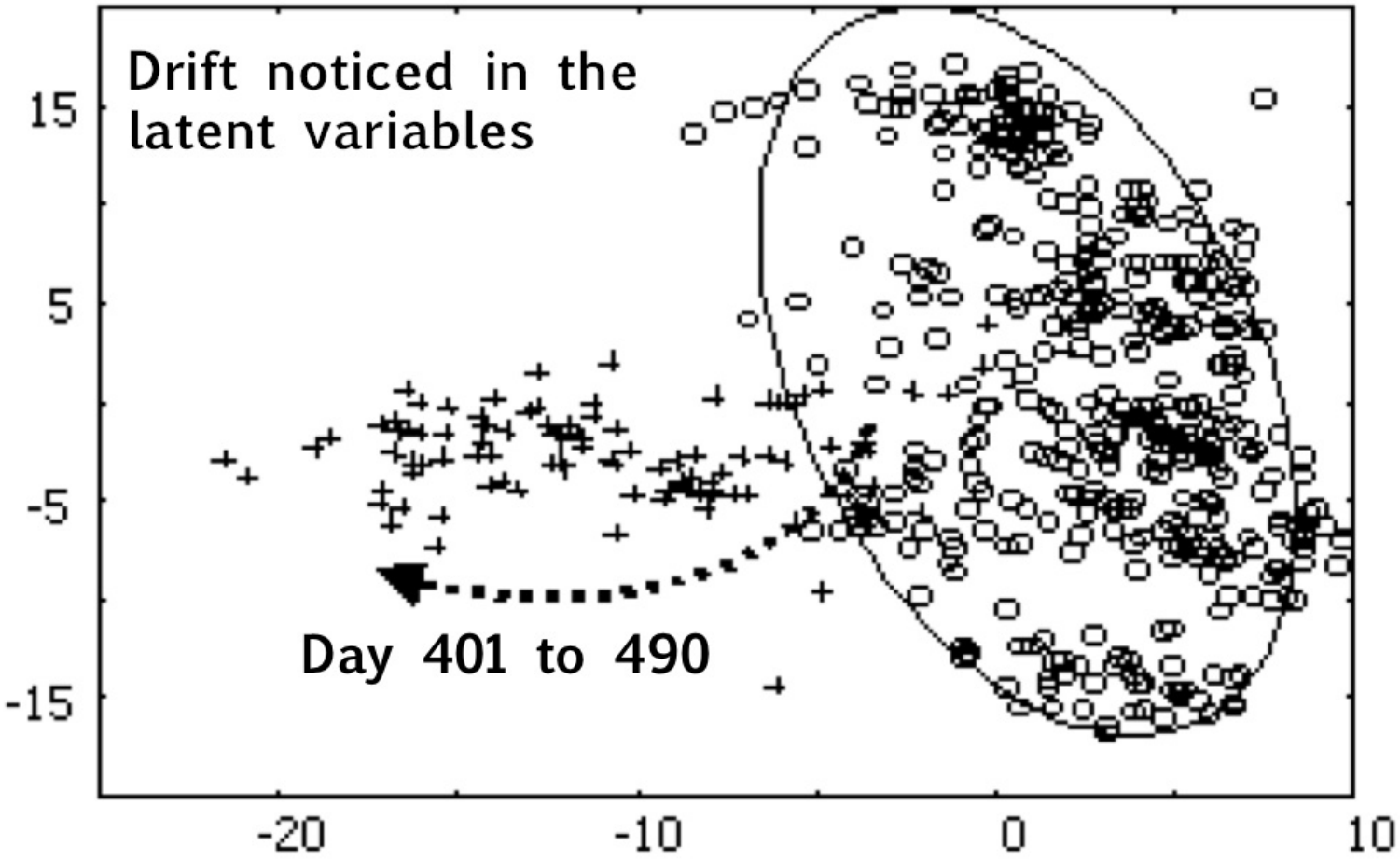
Process monitoring with PCA: SPE

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

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



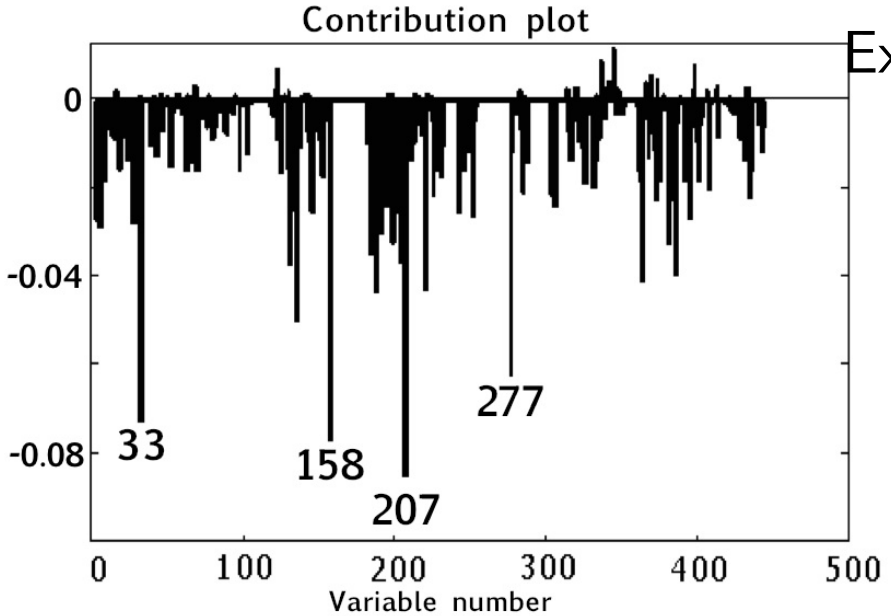
Diagnosing a problem



- 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

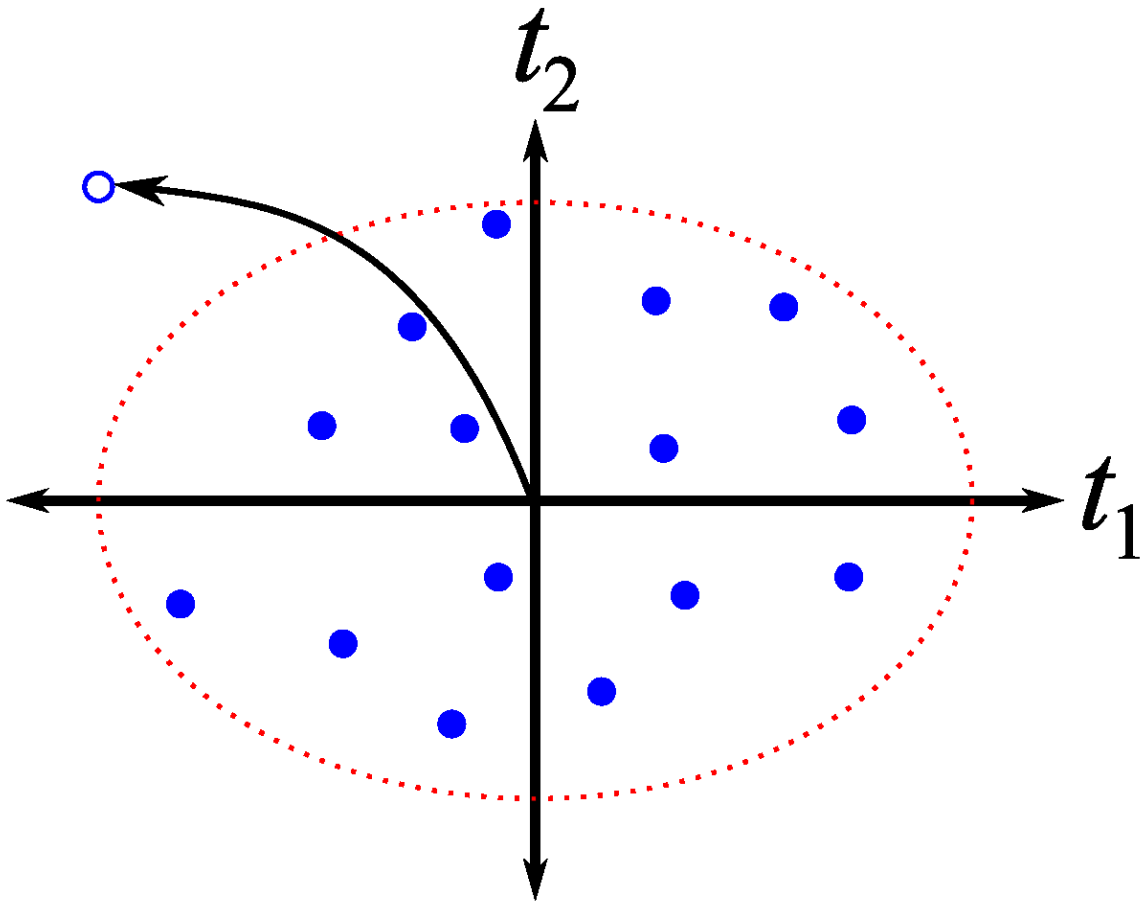
- These variables are related to the problem
- *Not the cause of the problem*
- Still have to use your engineering judgement to diagnose
- But, we've reduced the size of the problem

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*

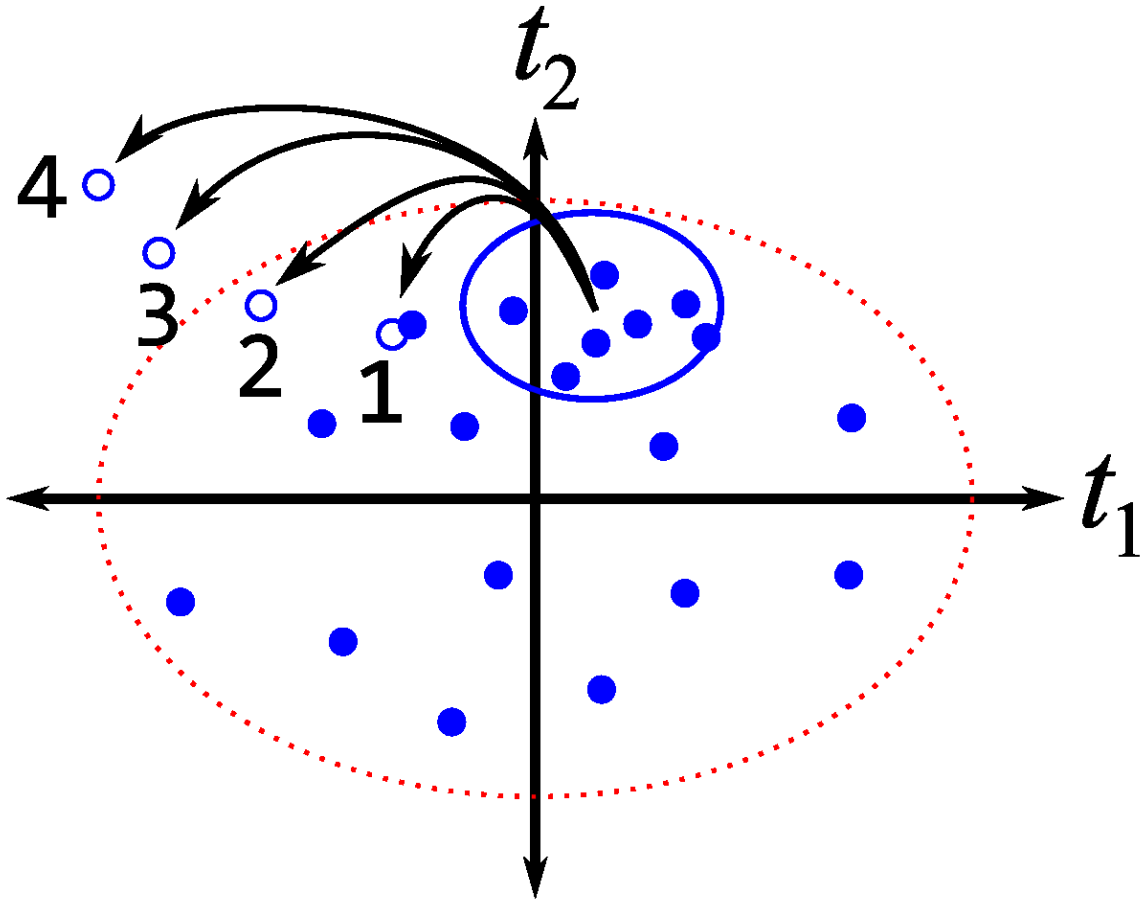
- T^2 contributions: weighted sum of scores
  - More details in Alvarez *et al.* - paper 21
  - and Kourti and MacGregor - paper 81

Contributions in the score space



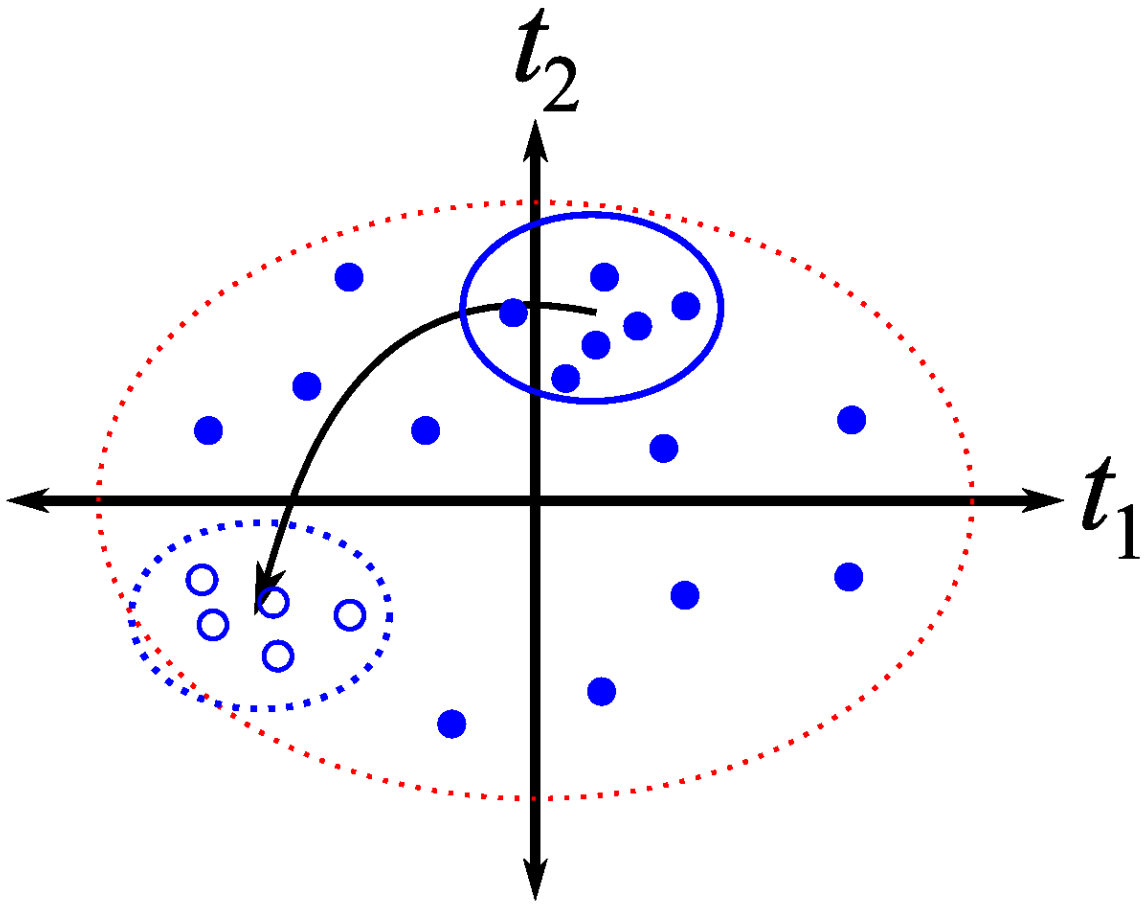
From the model center to a point

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

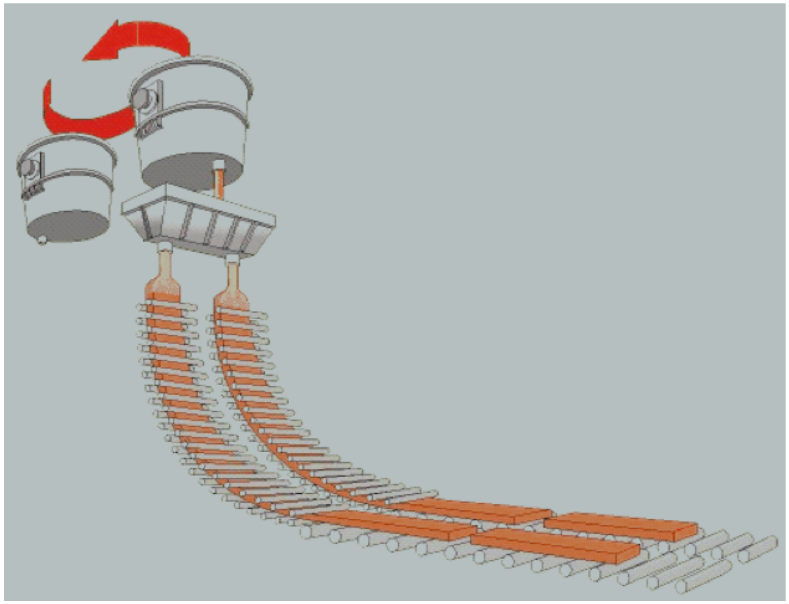
Industrial case study: Dofasco

- 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

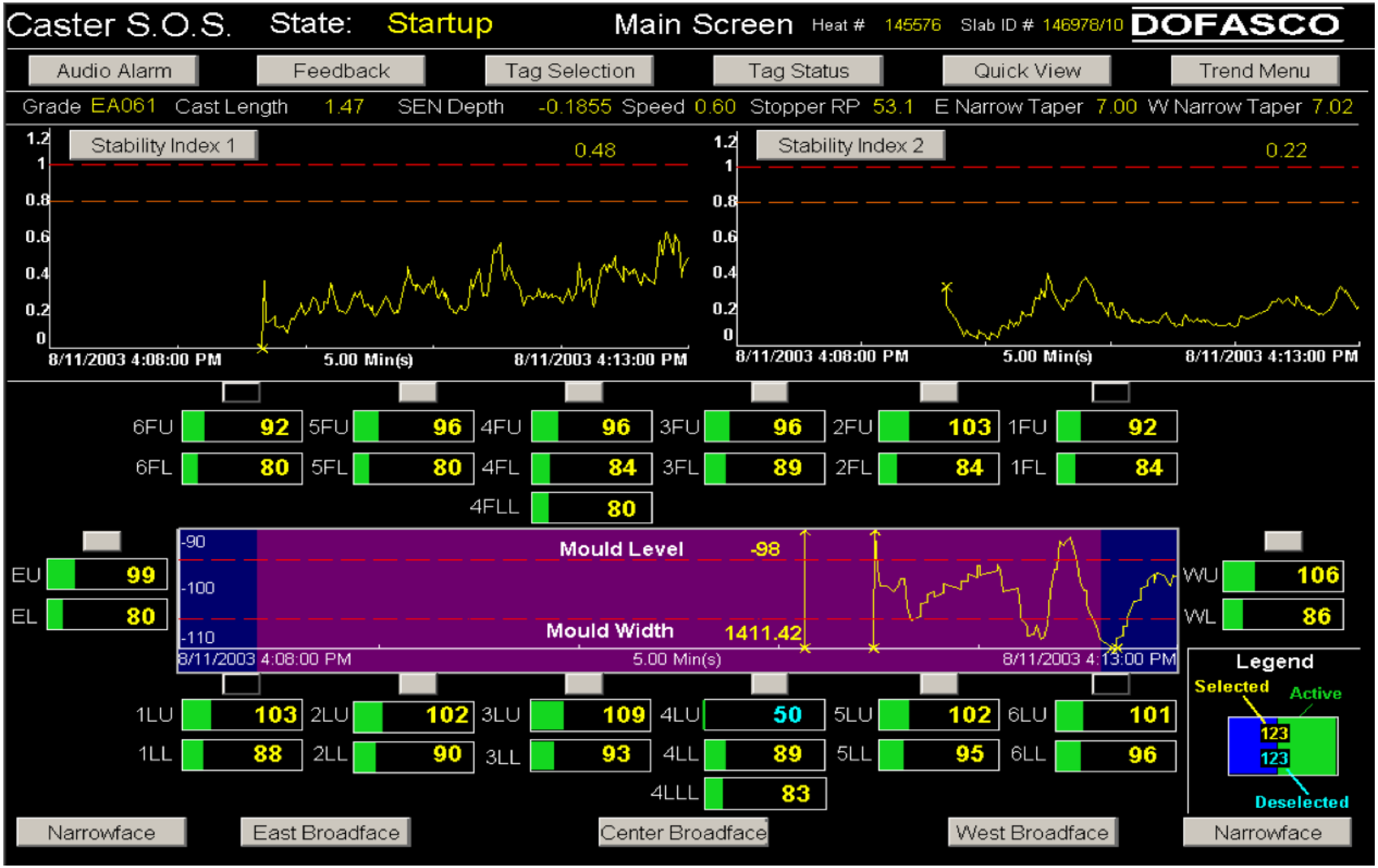
All screenshots with permission of Dr. John MacGregor

Dofasco case study: casting

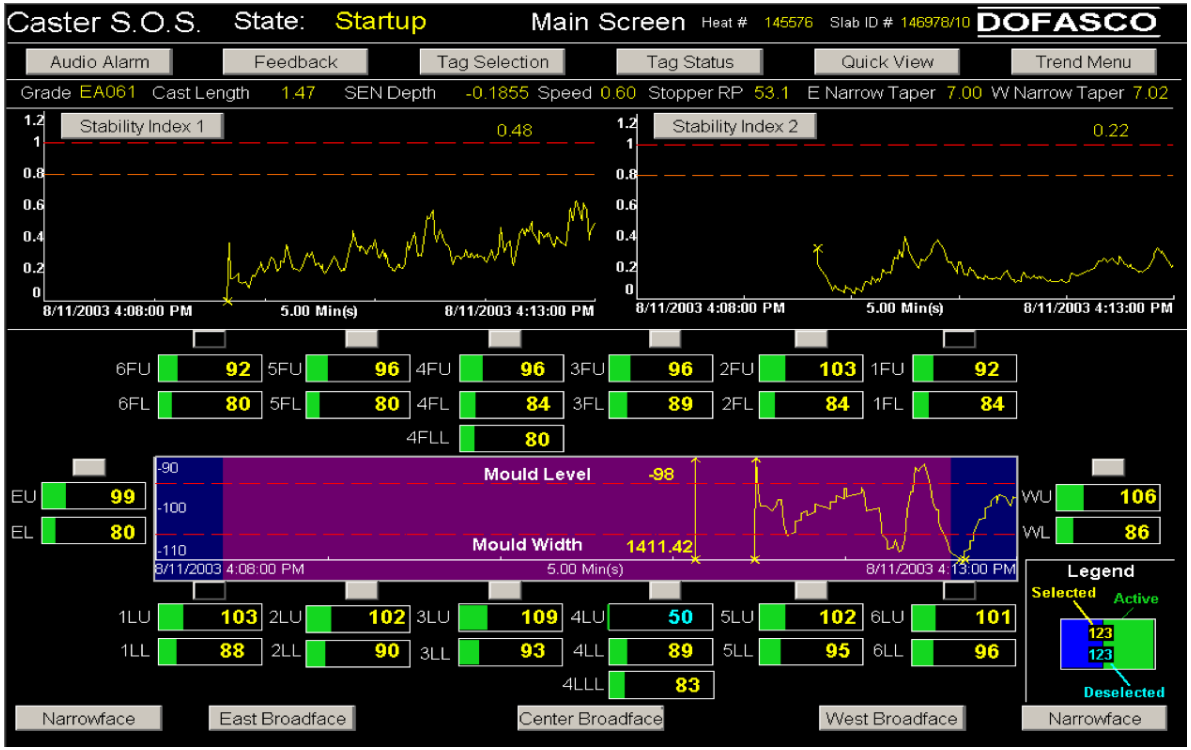


Dofasco case study: breakout

Dofasco case study: monitoring for breakouts

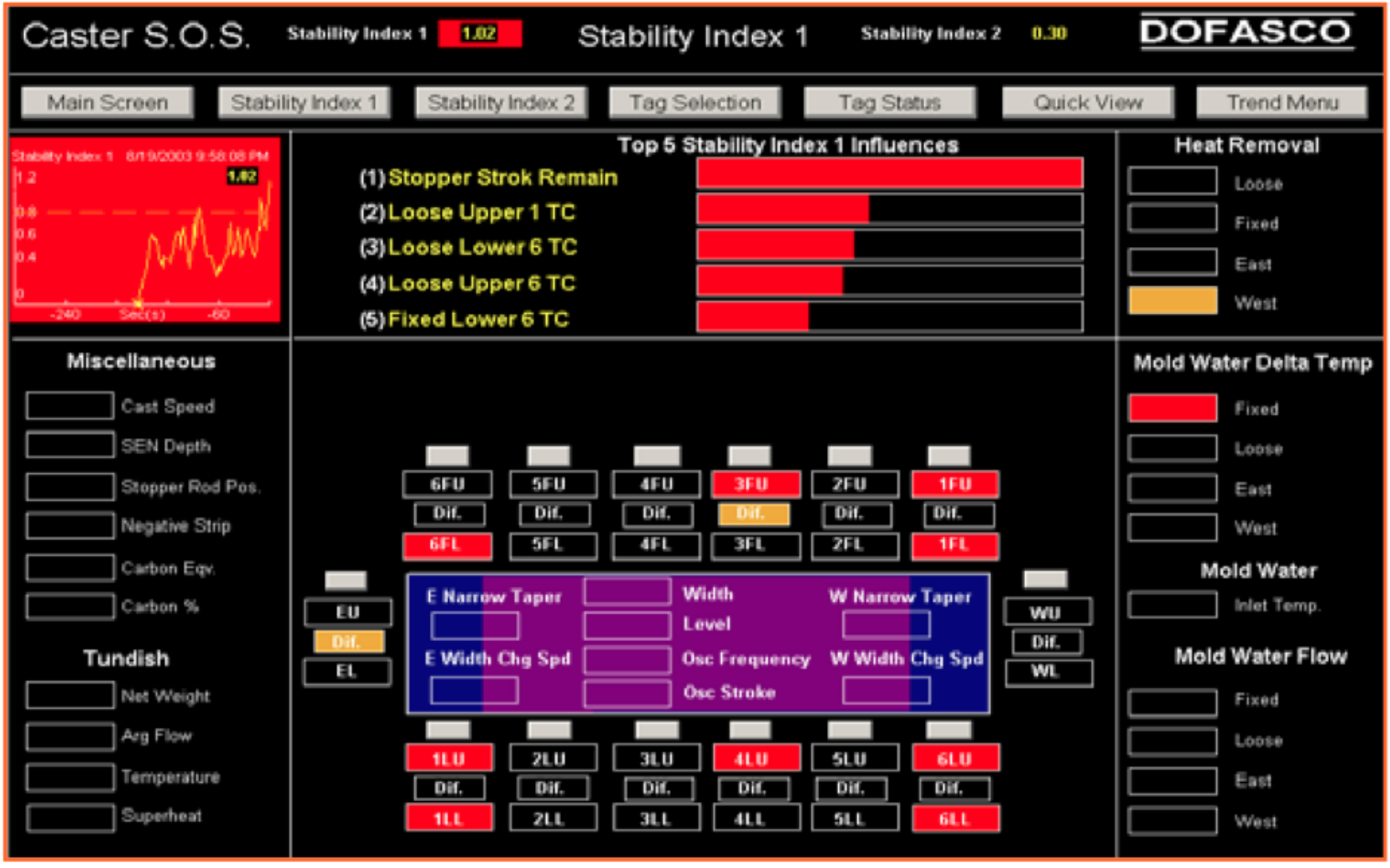


Dofasco case study: monitoring for breakouts



- 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

Dofasco case study: an alarm



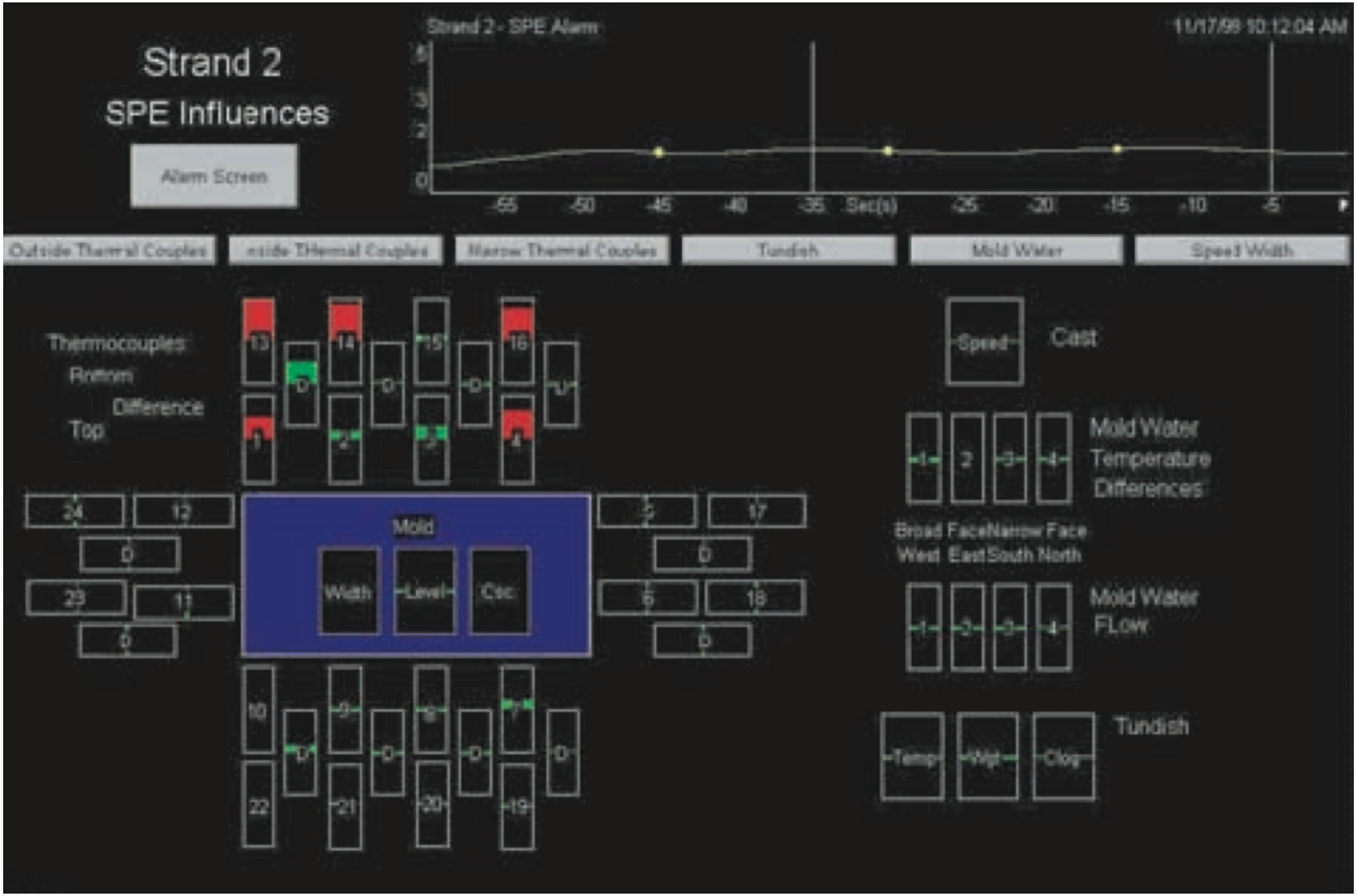
Dofasco case study: previous version

A previous version of the monitoring chart:

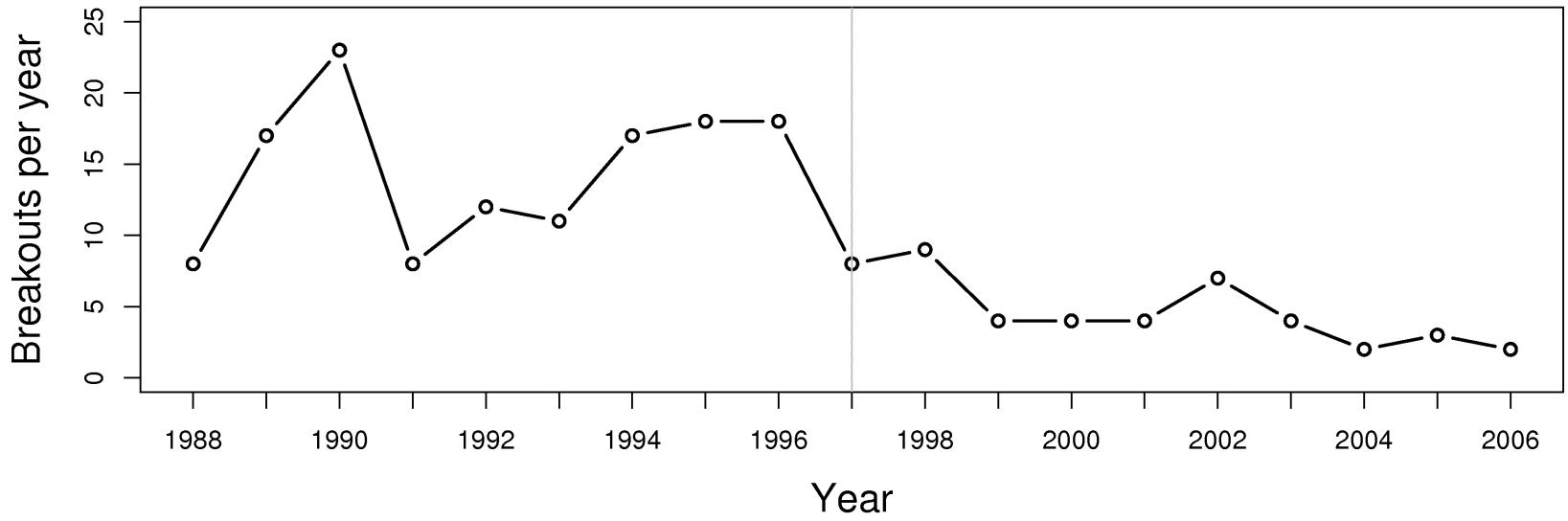
Updated based on operator feedback/requests

Dofasco case study: contribution plots

Contribution plot (previous version):

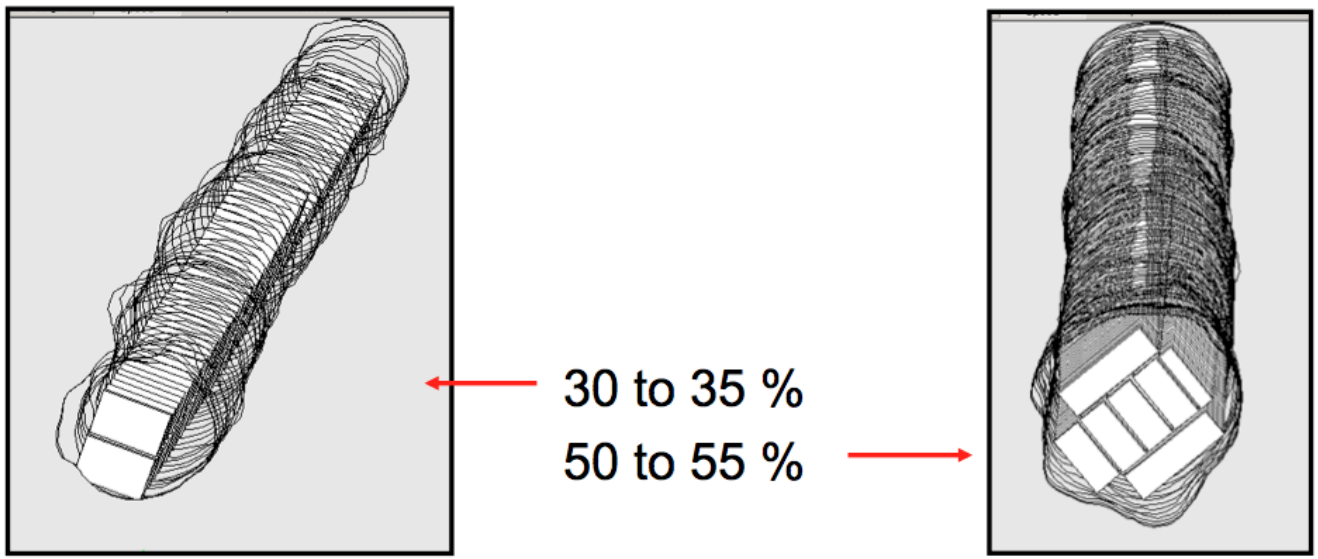


Dofasco case study: economics of monitoring



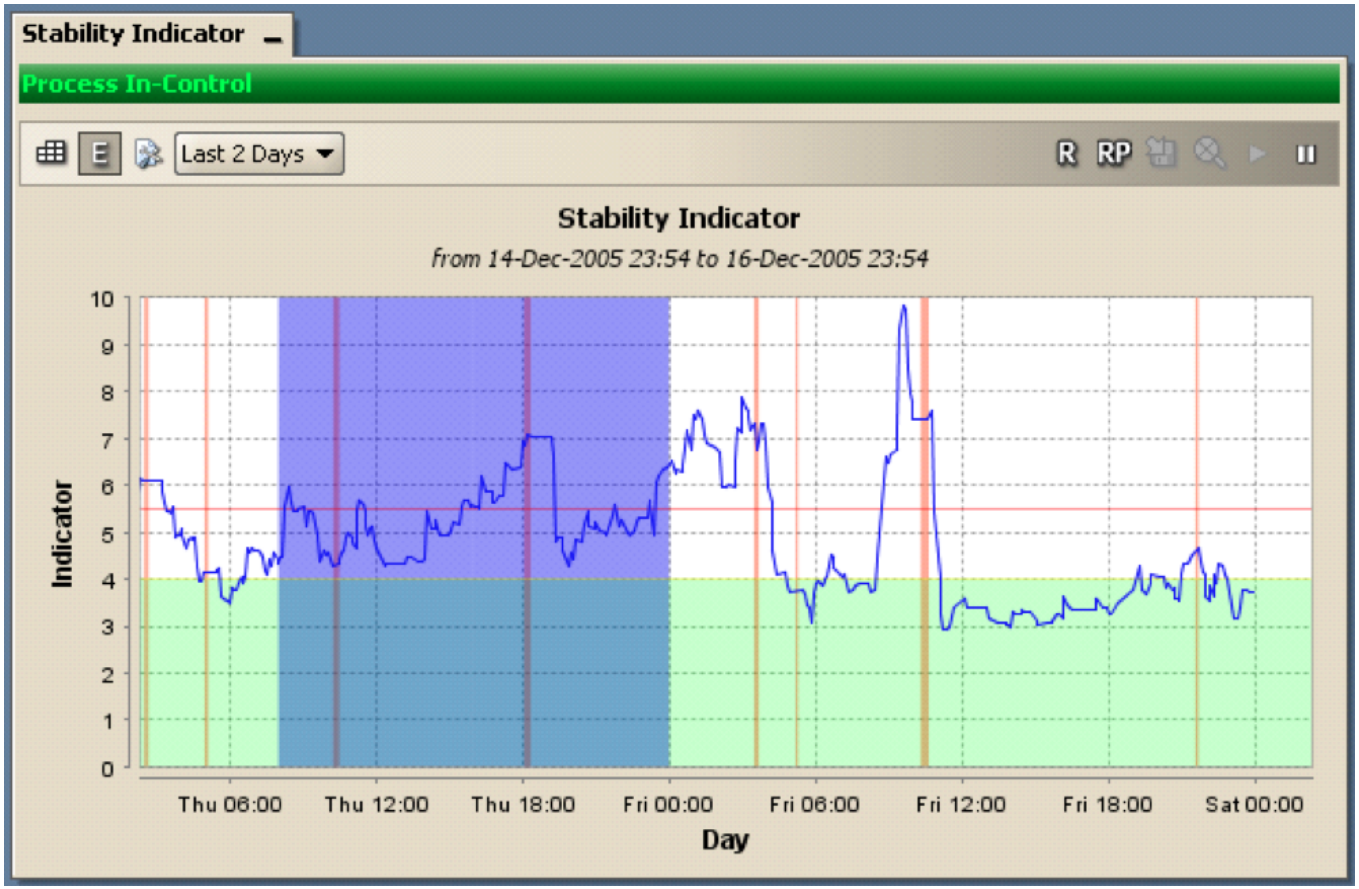
- Implemented system in 1997; multiple upgrades since then
- Economic savings: more than $ 1 million/year
  - each breakout costs around $200,000 to $500,000
  - process shutdowns and/or equipment damage

Lumber case study



Show video

Lumber case study



- Hotelling's T^2 is called "stability indicator" for operators
- Horizontal red line is the 99% limit
- Shaded green area is the 0 to 95% limit region

Monitoring isn't just for chemical processes

Any data stream can be monitored

- Raw material characteristics
- On-line data from systems (most common multivariate monitoring)
- Final quality properties
- End-point detection
- More generally: **any row in a data matrix**
  - Credit card/financial fraud monitoring
  - Human resources

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

Challenges for real-time monitoring

- Getting the data out
- Real-time use of the data (value of data decays exponentially)
- Training people to use the monitoring system is time consuming
- Bandwidth/network/storage/computing

Important readings

These papers will help you get to the bottom of process monitoring:

- MacGregor: Using on-line process data to improve quality: challenges for statisticians (paper 75)
- Kourti and MacGregor: Process analysis, monitoring and diagnosis, using multivariate projection methods (paper 31)
- MacGregor and Kourti: Statistical process control of multivariate processes (paper 16)
- Kresta, MacGregor and Marlin: Multivariate statistical monitoring of process operating performance (paper 9)
- Miller *et al.*: Contribution plots: a missing link in multivariate quality control (paper 78)