Linköping University
INSTITUTE OF TECHNOLOGY

PIC-LI

Data Mining from Historic Data for Process Identification

Daniel Peretzki[a,d], Alf J. Isaksson[a,b],
André Carvalho Bittencourt[a], Krister Forsman[c]

[a]Linköping University, Sweden

[b]ABB AB, Sweden

[c]Perstorp AB, Sweden

[d]Exchange student from University of Kassel, Germany

Nordic Process Control Workshop 26 January, 2012

LiU expanding reality

Motivation

- **Models are important for**:
  - Controller tuning, Performance monitoring, optimization

- **Traditional methods for process models**:
  - Modeling from physical principles
  - Performing experiments

• Chemical plants are complex

• Time-consuming

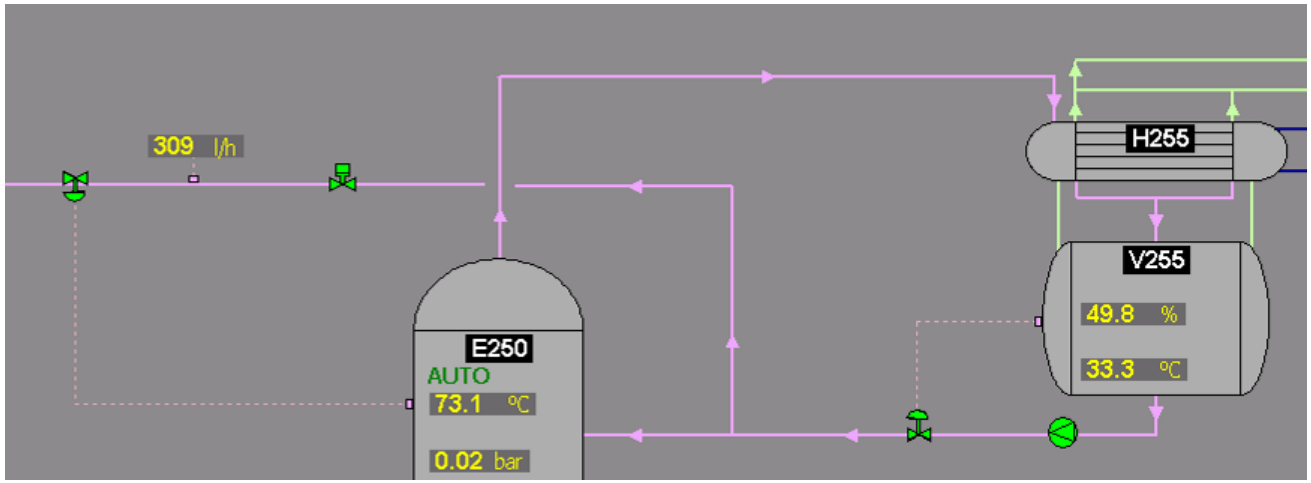
• May not be allowed due to risk of disturbing the process

- **Huge amounts of data**:
  - Continuously log data from each control loop
  - Data is stored indefinitely in a database

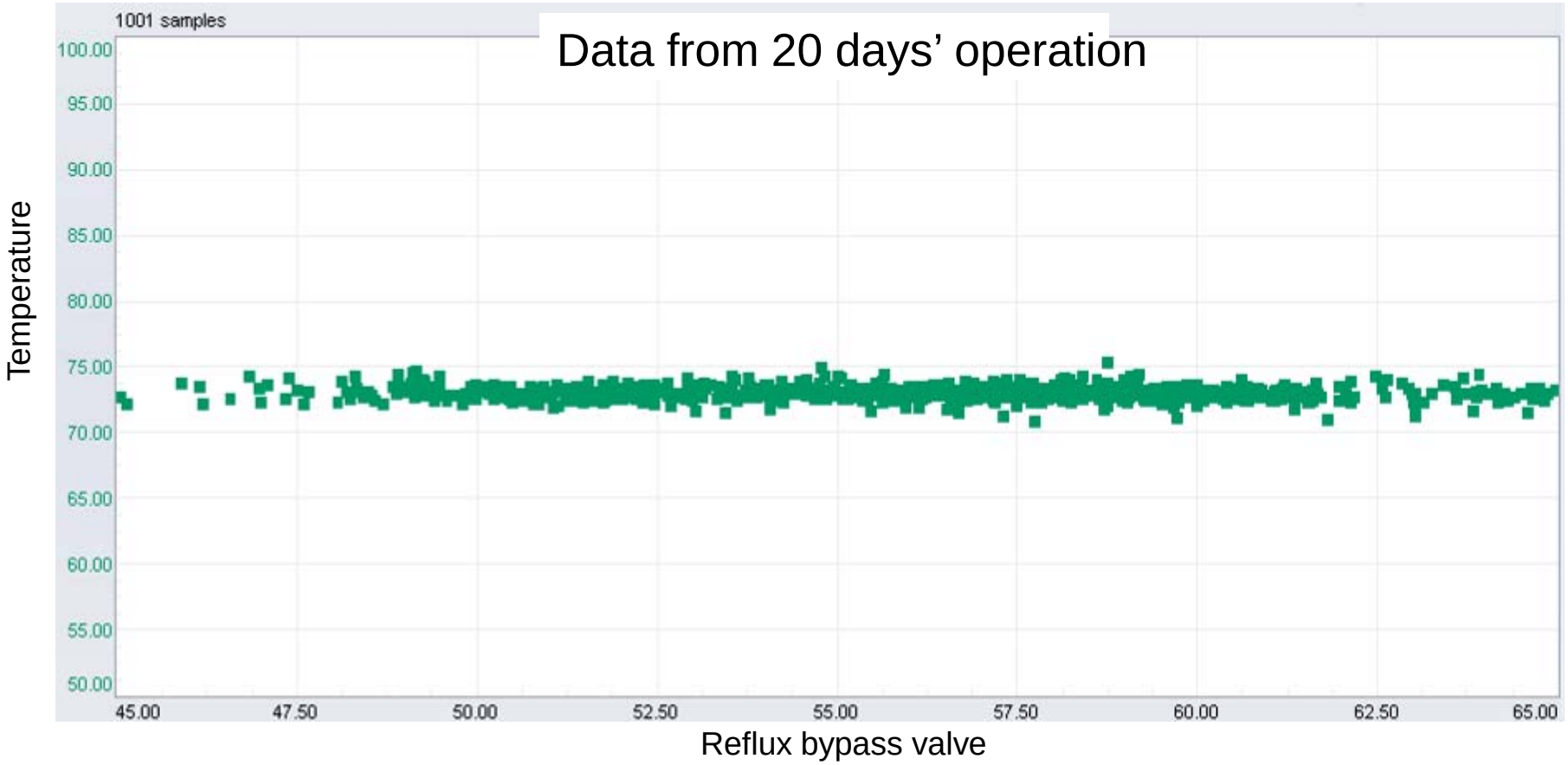
- **Cooperation with Perstorp**
  - Specialty Chemicals company
  - Provided 3 years' of data for 211 control loops

Danger of data tunnel vision; Example

- We wish to analyze a distillation column, using data from normal operations.
- What is the relation between temperature and reflux in the actual process? (Not in theory)
- Simplest idea: Collect data and make statistical analysis



What conclusion do we draw from data?



Watch out when drawing conclusions

- ”Reflux doesn’t affect temperature”
  - False
- ”There is no correlation between reflux and temperature, in the data.”
  - True, but useless
- ”There are other variations affecting temperature, that the controller compensates for by modifying reflux.”
  - Better description of the real world process
- Important conclusion:

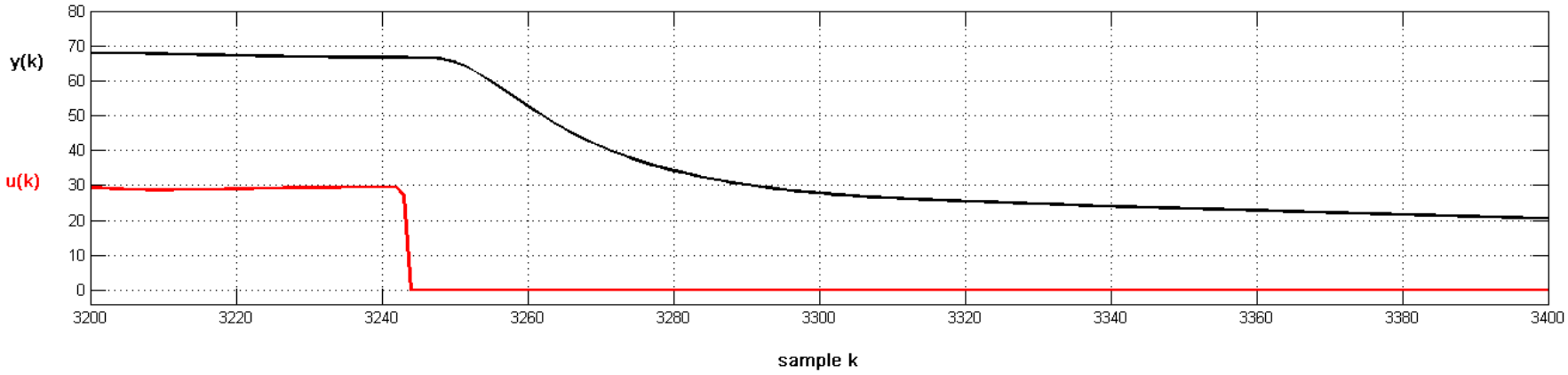
Data collected from a controlled process can normally not be used for modeling the process!

Data features for system identification

To estimate a process model we need:

- Excitation of:
  - the controller output *u(k)* when in manual mode (open loop)
  - the setpoint *r(k)* when in automatic mode (closed loop)
- Significant response in the process value *y(k)*
- A good signal-to-noise ratio

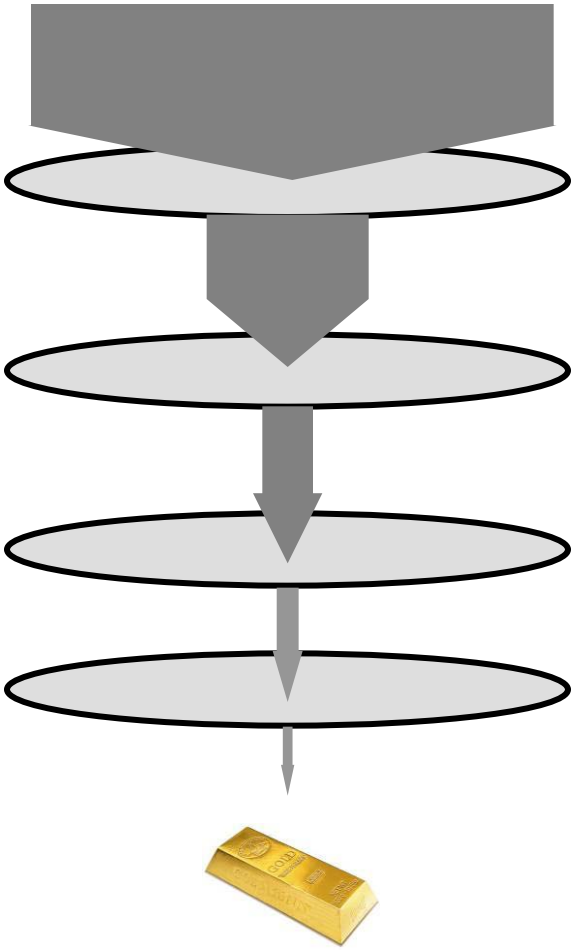
Real example (manual mode):



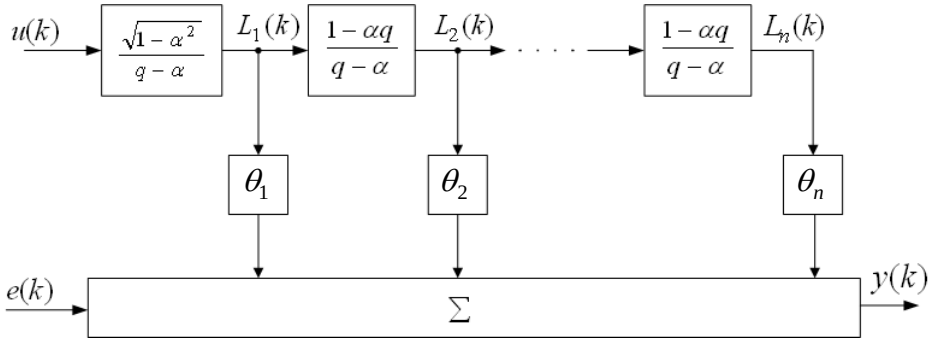
The method: ”a series of sieves”

Step-wise method to reduce computational burden. For each loop:

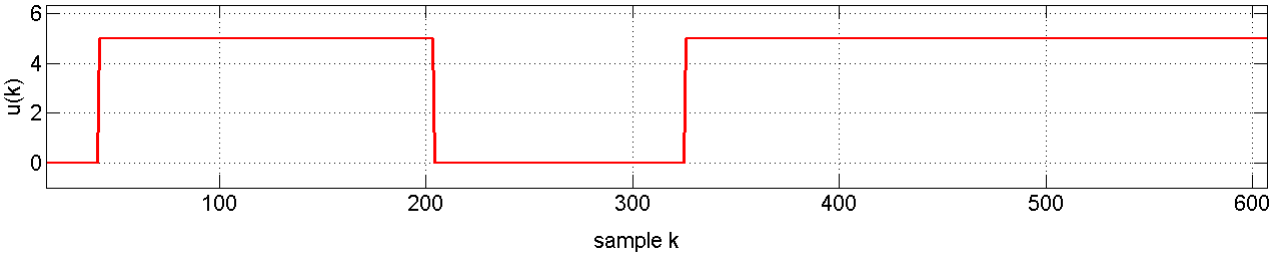
1. Partition data after controller operational mode
2. Scan for variation in *u*
3. Scan for significant variance in *y*
4. Scan for low condition number of information matrix
5. Estimate models, and verify parameter significance

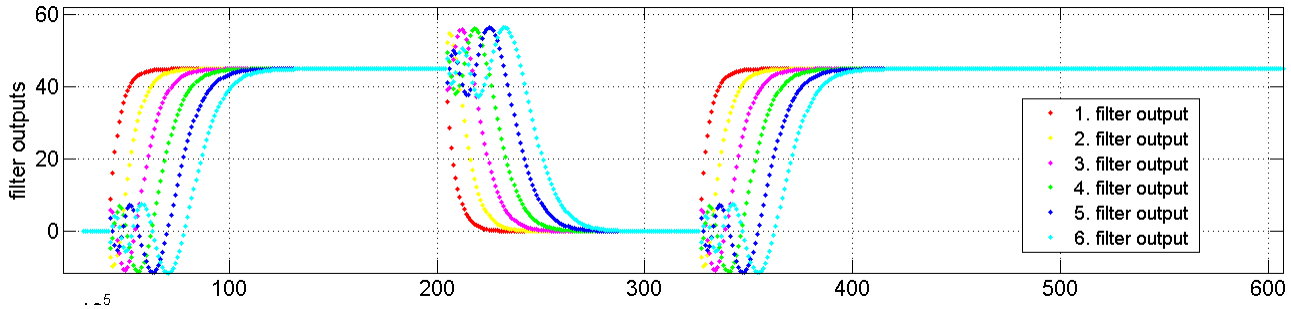


Remark: Laguerre model key component



The filters approximate first order plus time delay systems





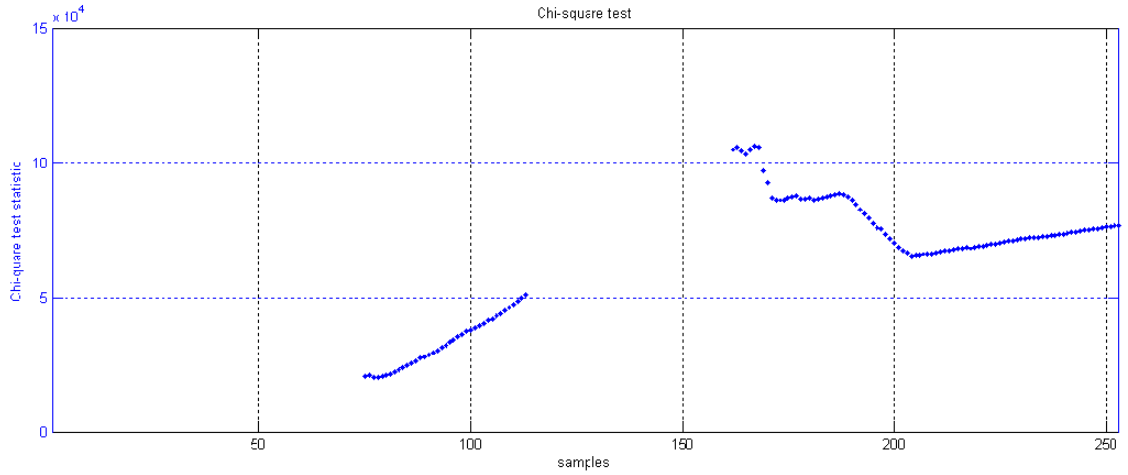
Method description; Step 5

Step 5: Chi-square test (check if any parameter is non-zero). Calculate

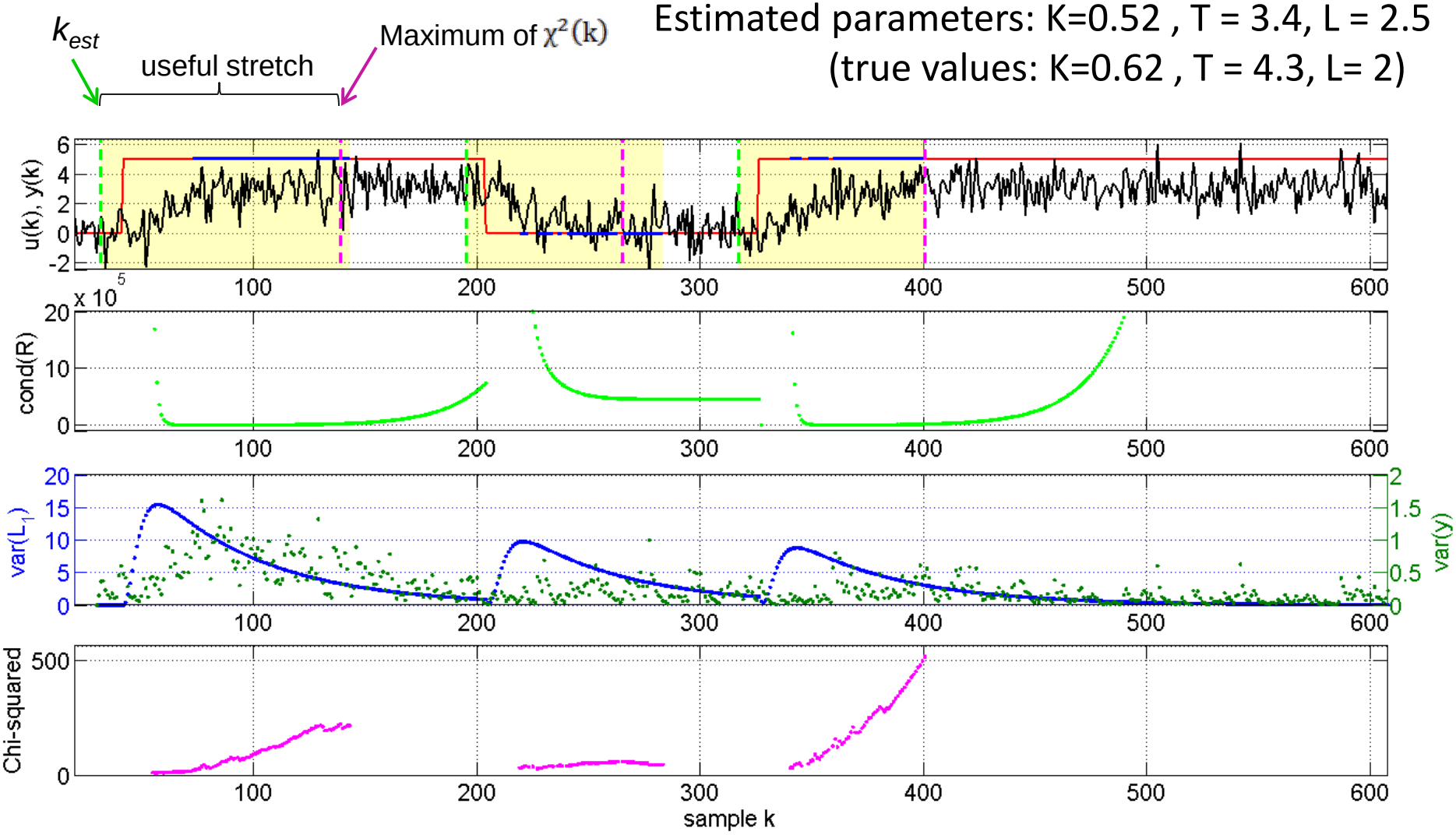
$$\chi^2(\mathrm{k}) = \hat{\theta}^T(k) \cdot \mathrm{P}(\mathrm{k})^{-1} \cdot \hat{\theta}(k)$$

If $\chi^2 > th_{conf}$ then parameters are significant enough and interval is marked as useful

Go to next sample $k = k+1$ and repeat steps 2 - 5



Simulated data example 1



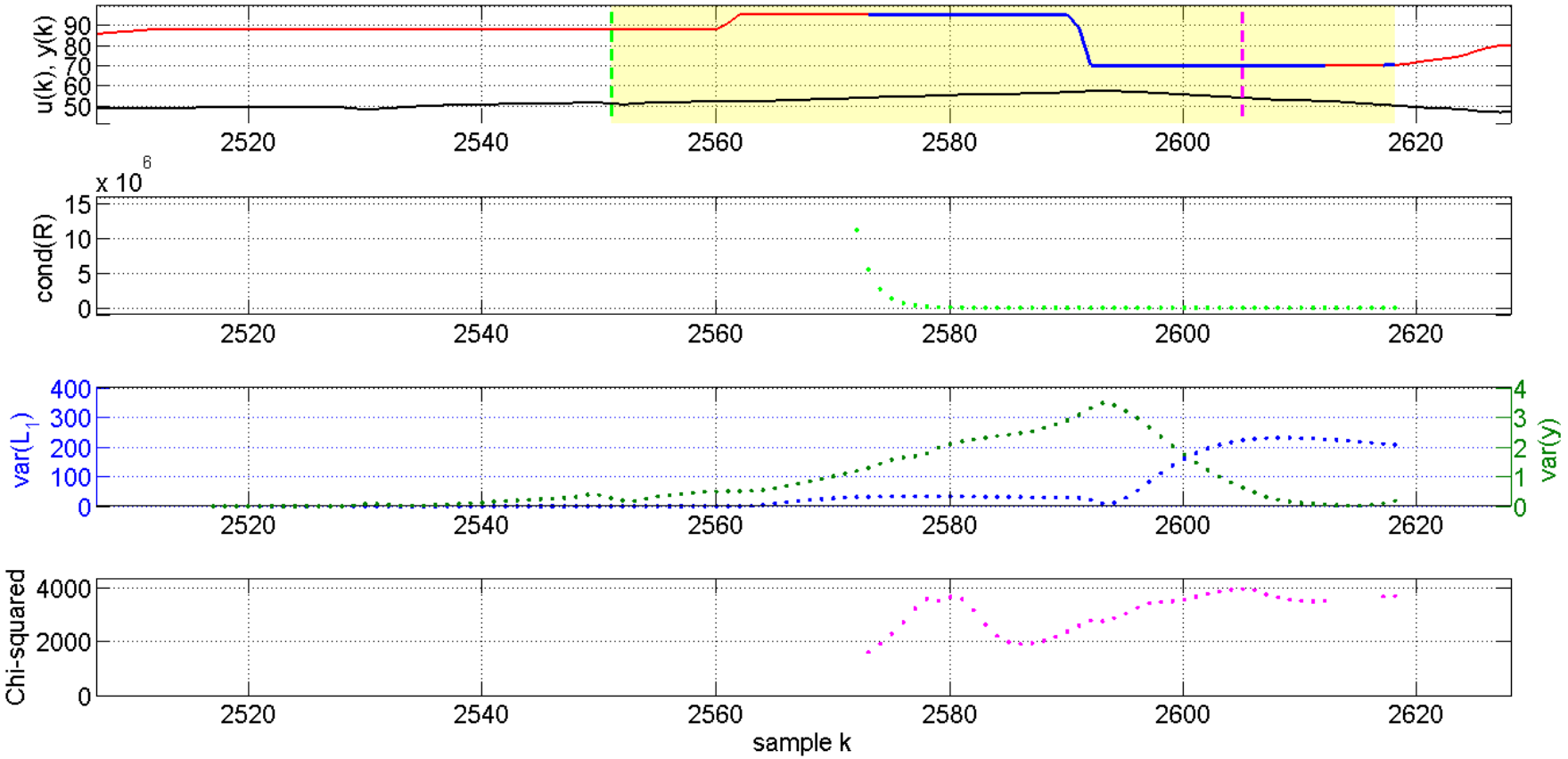
Estimated parameters: K=0.52 , T = 3.4, L = 2.5
(true values: K=0.62 , T = 4.3, L= 2)

Real data example 1

Known bump test (level control)

Estimated parameters: K=0.080, L= 1

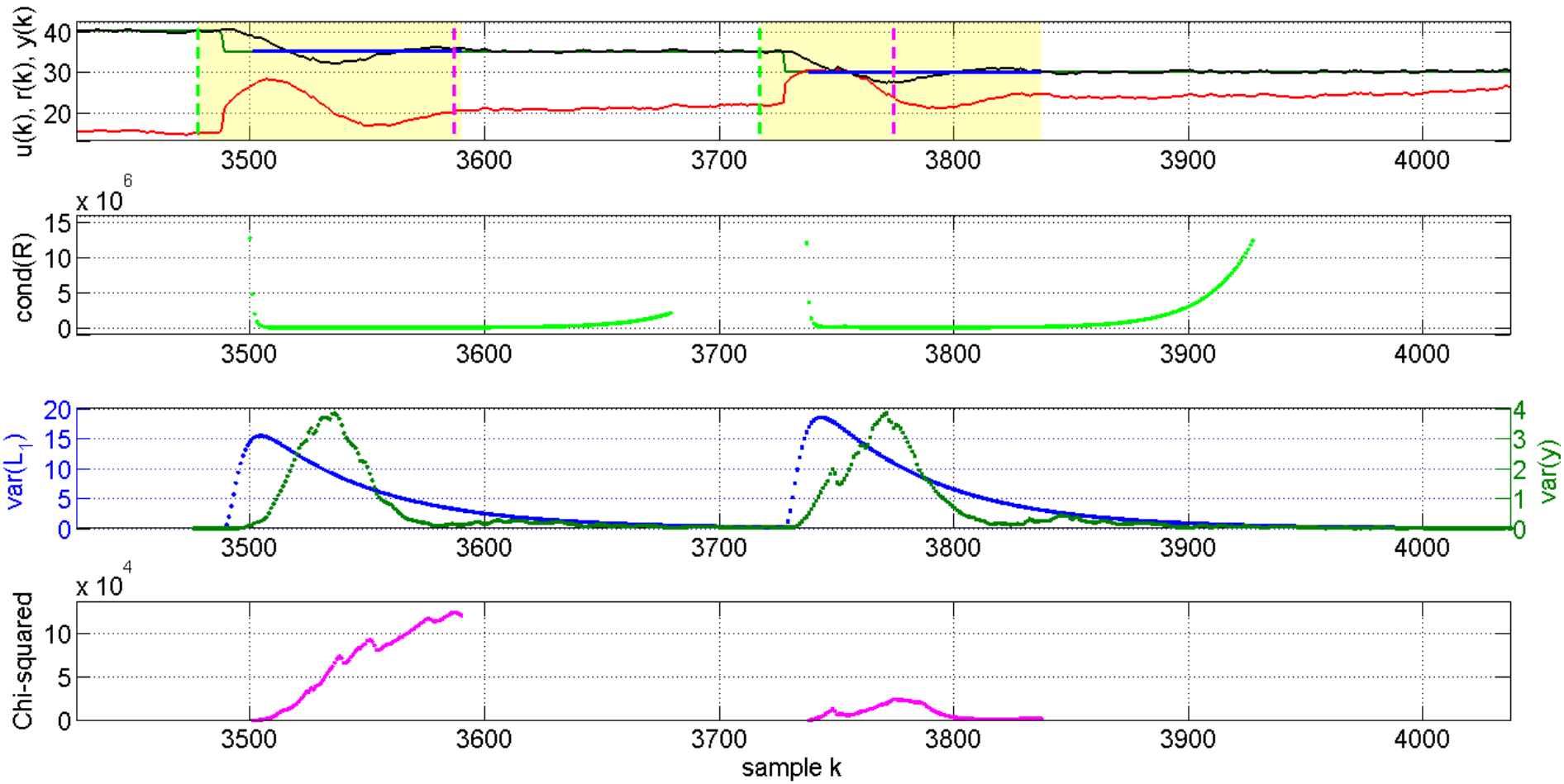
(Perstorp's result: K=0.083, L= 0)



Real data example 2

PIC-LI

Process in closed-loop (density)



Linköping University
INSTITUTE OF TECHNOLOGY

Verification of performance

3 years of data from 211 loops

- 1.1 G samples, 6.7 GB of data
- Computation time: ~ 1.5h
- ~ 1.5% of gold was found

Validation

- The method finds all intervals in the data where known identification experiments were performed
- Found additionally other useful intervals in closed/open loop operation

Loops where any Δ was found by mode	
closed loop	143 (67.7%)
open loop	185 (87.7%)
both	190 (90.1%)
Average length of Δ's found (samples) by mode	
closed loop	102.8
open loop	125.3
both	114.1
Average Δ's found by loop type	
Density	239
Flow	660
Concentration	84
Level	130
Conductivity	0
Temperature	35.3
Pressure	100

Summary and outlook

- Process identification by scanning huge data sets
- Essential features of the proposed search method:
  - Scan for excitation of the input and output (variances & condition number)
  - Estimate Laguerre model and check its significance by a chi-square test
- Performance:
  - Algorithm finds all known bump tests
  - Fast, without requiring much process knowledge

Future work:

- Speed up the algorithm (e.g. recursively updating covariance matrix $P(k)$)
- Accuracy of the method can be improved (fine tune design parameters)
- Test the method with data from other plants

Final remarks

- ABB has applied for a patent. Perstorp can use it freely in their plants.
- Presented at AIChE Annual Meeeting 2011
- Received prize for best MSc thesis by German process automation organization NAMUR.