Probabilistic mixtures of autoregressive models and their use in control

Václav Šmídl
(Pavel Ettler, Miroslav Kárný, Josef Andrýsek,...)

Institute of Information Theory and Automation,
Academy of Sciences,
Prague, Czech Republic

Workshop of WG on model-based clustering,
Dublin, Ireland

Outline

Background
- Adaptive Control
- Bayesian Adaptive Control
- Roles of mixture modelling

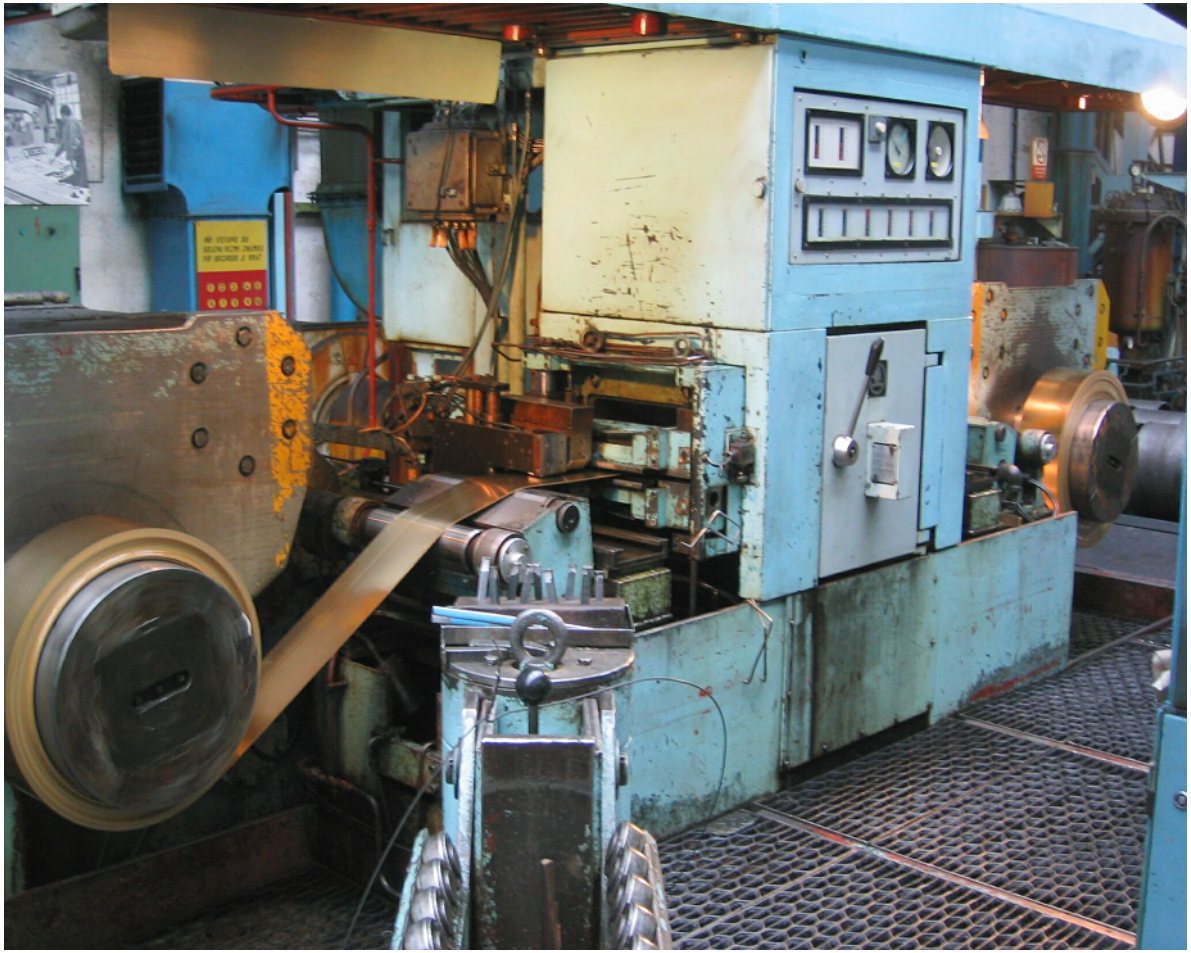
ProDaCTool project
- Operator control
- Recursive estimation of a mixture model
- Alternatives to EM
- List of off-line methods

Current and future work

Conclusion

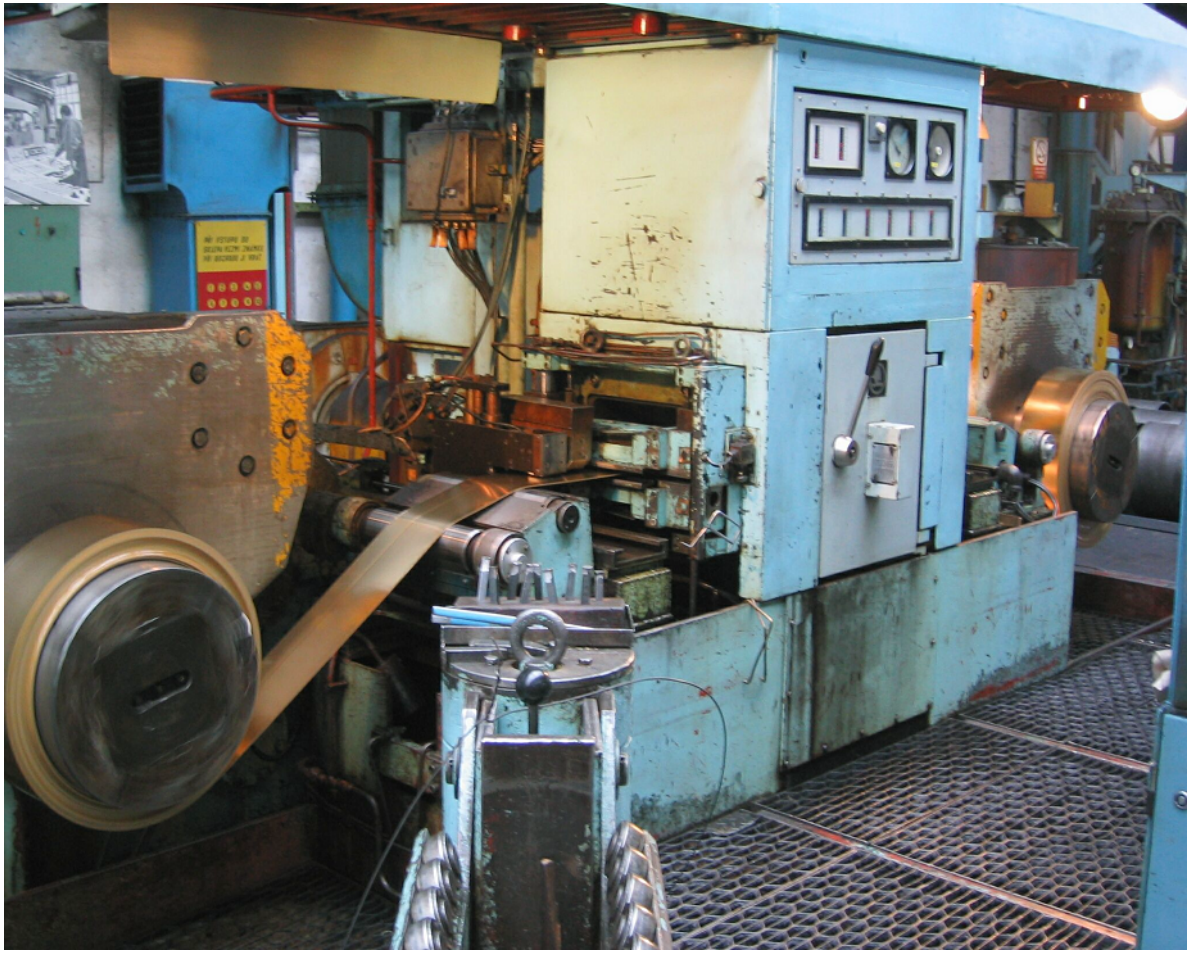
Data set? Data machine!

High-precision cold metal rolling mill
(Kovohute Rokycany)



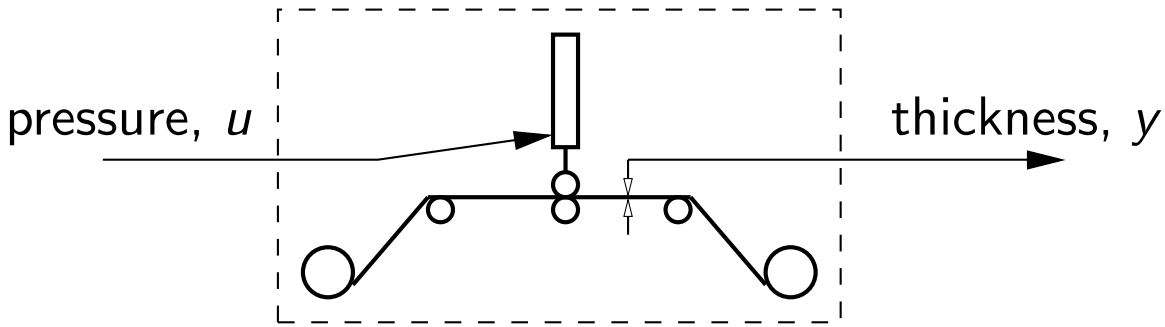
Data set? Data machine!

High-precision cold metal rolling mill
(Kovohute Rokycany)



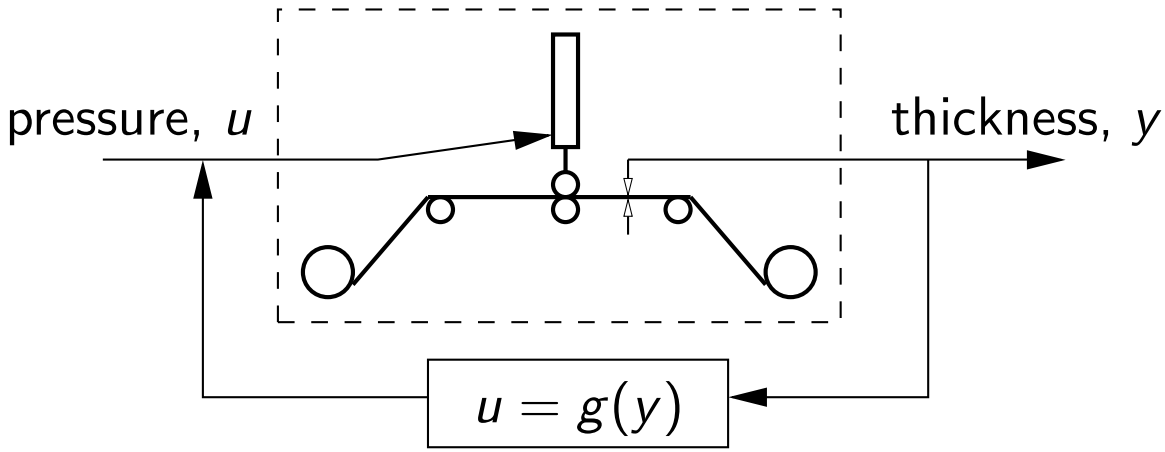
- 40 measured variables
- sampling time: milliseconds
- control action in each step
- GB of data and growing each day by MB

Feedback Control



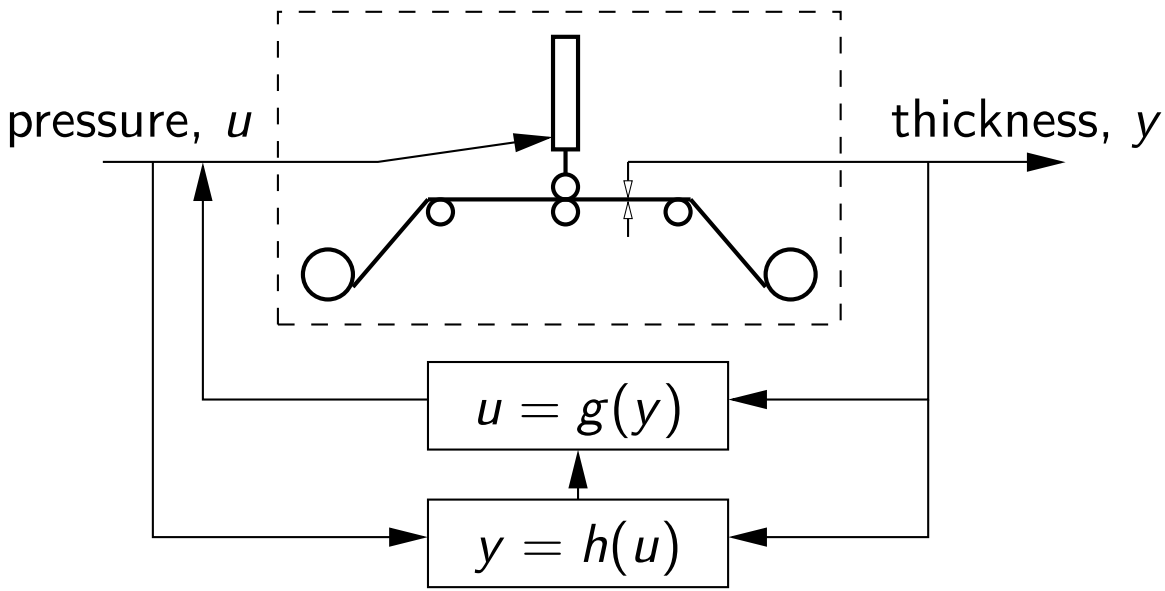
Negative feedback: if the output is too thick, increase pressure.

Feedback Control



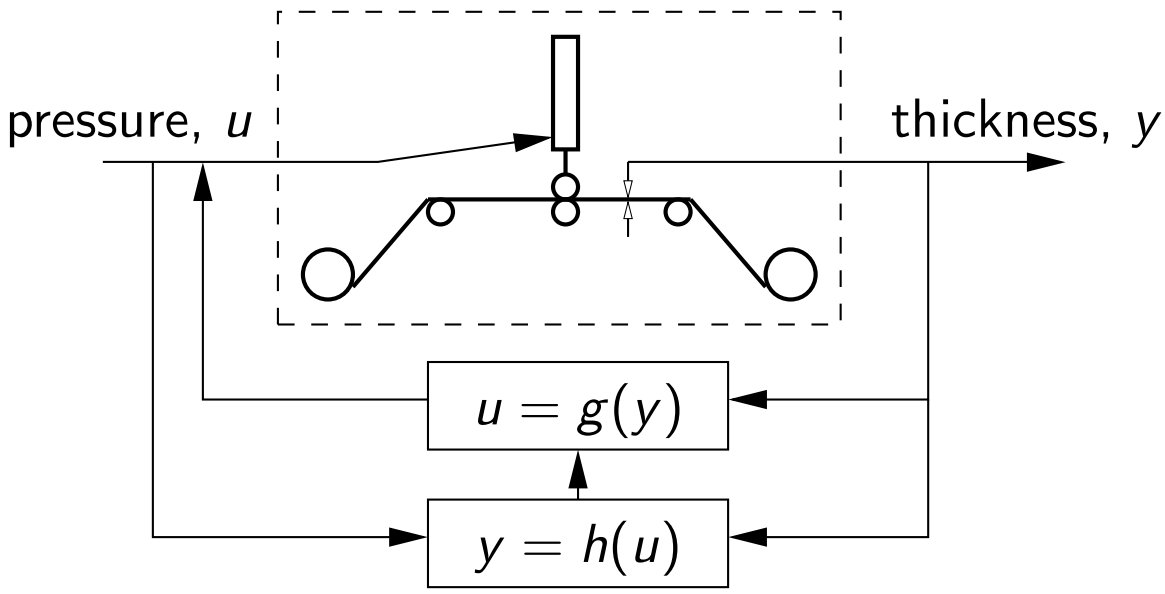
- Classical control is deterministic. $g()$ is designed using laws of physics.
- Problem with uncertainty:
  - material quality,
  - internal state of the machine,
  - sensors.
- Robust control: design $g()$ that performs well for a range of possible values.

Feedback Control



- Adaptive control:
  - model of the machine $y = h(u)$ is recursively estimated,
  - the control strategy is adapted *in each step*.
- Many possible approaches: gradient-based, neural network, AI, etc.

Feedback Control



- Adaptive control:
  - model of the machine $y = h(u)$ is recursively estimated,
  - the control strategy is adapted *in each step*.
- Many possible approaches: gradient-based, neural network, AI, etc.
- Bayesian identification, Peterka [1981].... and only a few of others:
  - Bayesian idea is too scary for control engineers,
  - Control systems are too complex for statisticians (optimality, consistency, etc.).

Why is estimation for control application different and difficult?

1. Nothing is i.i.d., when it is we have nothing to do.

Why is estimation for control application different and difficult?

1. Nothing is i.i.d., when it is we have nothing to do.
2. Restrictions: **recursivity**, infinite number of data, memory, computational time, etc.

Why is estimation for control application different and difficult?

1. Nothing is i.i.d., when it is we have nothing to do.
2. Restrictions: **recursivity**, infinite number of data, memory, computational time, etc.
3. Two very different systems (from control point of view):

$$\begin{aligned} y_t &= 0.95y_{t-1} - 2u_t - 2.05u_{t-1}, \\ y_t &= 0.95y_{t-1} - 2u_t - 1.95u_{t-1}. \end{aligned}$$

Why is estimation for control application different and difficult?

1. Nothing is i.i.d., when it is we have nothing to do.
2. Restrictions: **recursivity**, infinite number of data, memory, computational time, etc.
3. Two very different systems (from control point of view):

$$\begin{aligned} y_t &= 0.95y_{t-1} - 2u_t - 2.05u_{t-1}, \\ y_t &= 0.95y_{t-1} - 2u_t - 1.95u_{t-1}. \end{aligned}$$

4. **Feedback.** Changes model structure!

Example

ARX(2) model: $y_t = ay_{t-1} + bu_t + e_t.$

The aim is to approach $y_t \longrightarrow 0$.

Why is estimation for control application different and difficult?

1. Nothing is i.i.d., when it is we have nothing to do.
2. Restrictions: **recursivity**, infinite number of data, memory, computational time, etc.
3. Two very different systems (from control point of view):

$$\begin{aligned} y_t &= 0.95y_{t-1} - 2u_t - 2.05u_{t-1}, \\ y_t &= 0.95y_{t-1} - 2u_t - 1.95u_{t-1}. \end{aligned}$$

4. **Feedback.** Changes model structure!

Example

ARX(2) model: $y_t = ay_{t-1} + bu_t + e_t.$

The aim is to approach $y_t \longrightarrow 0$.

Why is estimation for control application different and difficult?

1. Nothing is i.i.d., when it is we have nothing to do.
2. Restrictions: **recursivity**, infinite number of data, memory, computational time, etc.
3. Two very different systems (from control point of view):

$$\begin{aligned} y_t &= 0.95y_{t-1} - 2u_t - 2.05u_{t-1}, \\ y_t &= 0.95y_{t-1} - 2u_t - 1.95u_{t-1}. \end{aligned}$$

4. **Feedback.** Changes model structure!

Example

ARX(2) model: $y_t = ay_{t-1} + bu_t + e_t$.
The aim is to approach $y_t \to 0$. Best control strategy is:

$$u_t = -\hat{a}/\hat{b}\, y_{t-1}$$

Why is estimation for control application different and difficult?

1. Nothing is i.i.d., when it is we have nothing to do.
2. Restrictions: **recursivity**, infinite number of data, memory, computational time, etc.
3. Two very different systems (from control point of view):

$$\begin{aligned} y_t &= 0.95y_{t-1} - 2u_t - 2.05u_{t-1}, \\ y_t &= 0.95y_{t-1} - 2u_t - 1.95u_{t-1}. \end{aligned}$$

4. **Feedback.** Changes model structure!

Example

ARX(2) model: $y_t = ay_{t-1} + bu_t + e_t$.
The aim is to approach $y_t \to 0$. Best control strategy is:

$$u_t = -\hat{a}/\hat{b}\, y_{t-1} \quad \Longrightarrow \quad y_t = (a - \hat{a}b/\hat{b})y_{t-1} + e_t.$$

Why is estimation for control application different and difficult?

1. Nothing is i.i.d., when it is we have nothing to do.
2. Restrictions: **recursivity**, infinite number of data, memory, computational time, etc.
3. Two very different systems (from control point of view):

$$\begin{aligned} y_t &= 0.95y_{t-1} - 2u_t - 2.05u_{t-1}, \\ y_t &= 0.95y_{t-1} - 2u_t - 1.95u_{t-1}. \end{aligned}$$

4. **Feedback.** Changes model structure!

Example

ARX(2) model: $y_t = ay_{t-1} + bu_t + e_t$.
The aim is to approach $y_t \to 0$. Best control strategy is:

$$u_t = -\hat{a}/\hat{b}\, y_{t-1} \quad \Longrightarrow \quad y_t = (a - \hat{a}b/\hat{b})y_{t-1} + e_t.$$

Dual control, Feldbaum, [1960], a dream of generations. Compromise between information richness and good control.

Bayesian Adaptive Control

Based on probabilistic *dynamic* model:

$$y_t \sim f(y_t|\psi_t, \theta).$$

vector ψ_t is composed of observations, e.g.
$\psi_t = [u_{t-20}, u_{t-21}, \ldots y_{t-1}, y_{t-1}, \ldots]$.

Off-line:

On-line:

Bayesian Adaptive Control

Based on probabilistic *dynamic* model:

$$y_t \sim f(y_t|\psi_t, \theta).$$

vector ψ_t is composed of observations, e.g.
$\psi_t = [u_{t-20}, u_{t-21}, \ldots y_{t-1}, y_{t-1}, \ldots]$.

Off-line: model structure, prior information $f_0(\theta)$,

On-line:

Bayesian Adaptive Control

Based on probabilistic *dynamic* model:

$$y_t \sim f(y_t|\psi_t, \theta).$$

vector ψ_t is composed of observations, e.g.
$\psi_t = [u_{t-20}, u_{t-21}, \ldots y_{t-1}, y_{t-1}, \ldots]$.

Off-line: model structure, prior information $f_0(\theta)$,

On-line:

1. update posterior parameter density

$$\begin{aligned} f(\theta|y_{(1:t)}, \psi_{(1:t)}) &\propto f(y_t|\psi_t, \theta) \times \ldots f(y_1|\psi_1, \theta) \times f_0(\theta) \\ &\propto f(y_t|\psi_t, \theta) \times f(\theta|y_{(1:t-1)}, \psi_{(1:t-1)}), \\ o_{(1:t)} &= [o_1, o_2, \ldots o_t]. \end{aligned}$$

2. find such control action u_{t+1} which minimizes expected future loss.

Recursive estimation

Evaluation of the Bayes rule for $t = 1, 2, \ldots, \infty$:

$$f(\theta|y_{(1:t)}, \psi_{(1:t)}) \propto f(y_t|\psi_t, \theta) f(\theta|y_{(1:t-1)}, \psi_{(1:t-1)}).$$

It is possible only if *finite-dimensional* sufficient statistics exist for all t:

$$f\left(\theta|y_{(1:t)}, \psi_{(1:t)}\right) = f\left(\theta|V_t\right).$$

This is guaranteed only within the exponential family:

$$f\left(y_t|\psi_t, \theta\right) = A(\theta) \exp\left(\langle B(y_t, \psi_t), C(\theta)\rangle + D(y_t, \psi_t)\right).$$

Then

$$f\left(\theta|V_t, \nu_t\right) \propto A^{\nu_t}(\theta) \exp\left(\langle V_t, C(\theta)\rangle\right)$$

$$V_t = V_{t-1} + B(y_t, \psi_t), \quad \nu_t = \nu_{t-1} + 1.$$

Recursive estimation

Evaluation of the Bayes rule for $t = 1, 2, \ldots, \infty$:

$$f(\theta|y_{(1:t)}, \psi_{(1:t)}) \propto f(y_t|\psi_t, \theta) f(\theta|y_{(1:t-1)}, \psi_{(1:t-1)}).$$

It is possible only if *finite-dimensional* sufficient statistics exist for all t:

$$f\left(\theta|y_{(1:t)}, \psi_{(1:t)}\right) = f\left(\theta|V_t\right).$$

This is guaranteed only within the exponential family:

$$f\left(y_t|\psi_t, \theta\right) = A(\theta) \exp\left(\langle B(y_t, \psi_t), C(\theta)\rangle + D(y_t, \psi_t)\right).$$

Then

$$f\left(\theta|V_t, \nu_t\right) \propto A^{\nu_t}(\theta) \exp\left(\langle V_t, C(\theta)\rangle\right)$$

$$V_t = V_{t-1} + B(y_t, \psi_t), \quad \nu_t = \nu_{t-1} + 1.$$

(Geometric approach, Kulhavý [1990]: estimating weights of a mixture.)

Limitations of EF

1. Few autoregressive members: Normal for continuous and Multinomial for discrete.
   Poisson distribution is in the exponential family:

$$f\left(y_t|\lambda\right) = \mathcal{P}o\left(\lambda\right) = \underbrace{\exp\left(-\lambda\right)}_{A} \exp\left(\underbrace{y_t}_{B} \underbrace{\log\left(\lambda\right)}_{C} \underbrace{-\log y_t!}_{D}\right),$$

Limitations of EF

1. Few autoregressive members: Normal for continuous and Multinomial for discrete.
   Poisson distribution is in the exponential family:

$$f\left(y_t|\lambda\right) = \mathcal{P}o\left(\lambda\right) = \underbrace{\exp\left(-\lambda\right)}_{A} \exp\left(\underbrace{y_t}_{B} \underbrace{\log\left(\lambda\right)}_{C} \underbrace{-\log y_t!}_{D}\right),$$

   2nd order auto-regressive Poisson distribution:

$$\begin{aligned} f\left(y_t|\theta, y_{t-1}, t_{t-2}\right) &= \mathcal{P}o\left(\theta_1 y_{t-1} + \theta_2 y_{t-2}\right) \\ &= \exp\left(-\theta_1 y_{t-1} - \theta_2 y_{t-2}\right) \times \\ &\quad \exp\left(y_t \left(\log\left(\theta_1 y_{t-1} + \theta_2 y_{t-2}\right)\right) - \log y_t!\right), \end{aligned}$$

Limitations of EF

1. Few autoregressive members: Normal for continuous and Multinomial for discrete.
   Poisson distribution is in the exponential family:

$$f\left(y_t|\lambda\right) = \mathcal{P}o\left(\lambda\right) = \underbrace{\exp\left(-\lambda\right)}_{A} \exp\left(\underbrace{y_t}_{B} \underbrace{\log\left(\lambda\right)}_{C} \underbrace{-\log y_t!}_{D}\right),$$

   2nd order auto-regressive Poisson distribution:

$$\begin{aligned} f\left(y_t|\theta, y_{t-1}, t_{t-2}\right) &= \mathcal{P}o\left(\theta_1 y_{t-1} + \theta_2 y_{t-2}\right) \\ &= \exp\left(-\theta_1 y_{t-1} - \theta_2 y_{t-2}\right) \times \\ &\quad \exp\left(y_t \left(\log\left(\theta_1 y_{t-1} + \theta_2 y_{t-2}\right)\right) - \log y_t!\right), \end{aligned}$$

2. Assumption of time-invariant parameters.

Time-varying parameters

Approaches:

1. Windowing, Jazwinski [1979]. Estimation is performed only on the last h data.
   - ⊕ allows to use off-line estimation techniques.
   - ⊖ could introduce time-delay (e.g. median estimation). (The worst artefact in control).

Time-varying parameters

Approaches:

1. Windowing, Jazwinski [1979]. Estimation is performed only on the last h data.
   - ⊕ allows to use off-line estimation techniques.
   - ⊖ could introduce time-delay (e.g. median estimation). (The worst artefact in control).
2. Forgetting. Bayesian interpretation, Kulhavý and Zarrop [1993], as a projection from two possible hypothesis into EF.
   - ⊕ has the same algebraic form as updates in EF,
   - ⊖ difficulties in finding alternative hypotheses.

Time-varying parameters

Approaches:

1. Windowing, Jazwinski [1979]. Estimation is performed only on the last h data.
   - ⊕ allows to use off-line estimation techniques.
   - ⊖ could introduce time-delay (e.g. median estimation). (The worst artefact in control).
2. Forgetting. Bayesian interpretation, Kulhavý and Zarrop [1993], as a projection from two possible hypothesis into EF.
   - ⊕ has the same algebraic form as updates in EF,
   - ⊖ difficulties in finding alternative hypotheses.
3. Bayesian filtering, Doucet et al. [2001]. Extension via parameter evolution model, $f(\theta_t|\theta_{t-1})$
   - ⊕ accuracy,
   - ⊖ computationally expensive, difficulties in designing the model.

Why mixtures?

1. Possible extension of EF
   - universal approximation property,
2. Higher level control, operator control.
   - not so time critical,
   - multi-modality

Why mixtures?

1. Possible extension of EF
   - universal approximation property,
2. Higher level control, operator control.
   - not so time critical,
   - multi-modality

Common interest with model-based clustering:

- Prior elicitation,
- Dimensionality reduction,
- Variable selection.
- Clustering (fault detection) and density estimation (modeling).

Why mixtures?

1. Possible extension of EF
   - universal approximation property,
2. Higher level control, operator control.
   - not so time critical,
   - multi-modality

Common interest with model-based clustering:

- Prior elicitation,
- Dimensionality reduction,
- Variable selection.
- Clustering (fault detection) and density estimation (modeling).

Differences:

1. Design of control strategy.
2. Recursive identification of non-i.i.d. models.

Why ARX models?

- experience: adaptive control of metal rolling mill, Ettler [1986],
- conjugate prior, analytical solution of recursive updates,
- (relatively) easy model structure determination (heuristic algorithm),
- computationally efficient and robust numerical algorithms via LD decompositions of $V_t = L_t' D_t L_t$,

$$L_t = \upsilon(L_{t-1}, y_t, \psi_t), \quad D_t = \upsilon(D_{t-1}, y_t, \psi_t),$$

- evaluation of predictive (marginal) probabilities of data,
- **reliable methods for design of control strategy (Riccatti equation).**
- Mixtures of ARX might be controllable by linear combination of Riccatti equations.

Outline

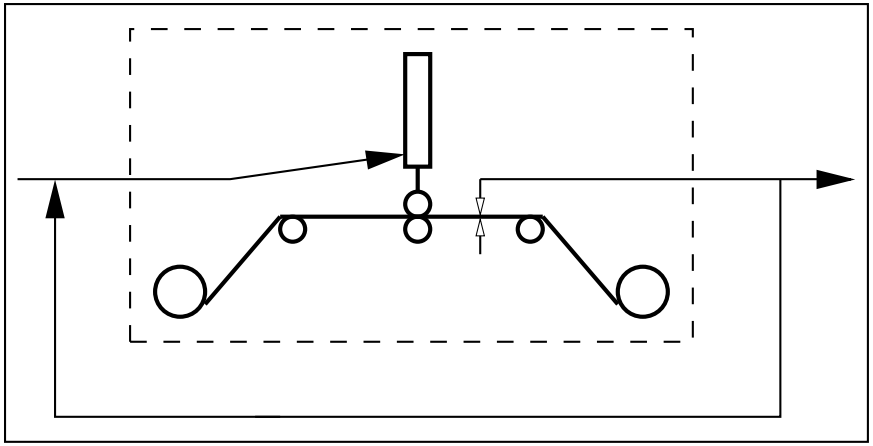
Background
- Adaptive Control
- Bayesian Adaptive Control
- Roles of mixture modelling

ProDaCTool project
- Operator control
- Recursive estimation of a mixture model
- Alternatives to EM
- List of off-line methods

Current and future work

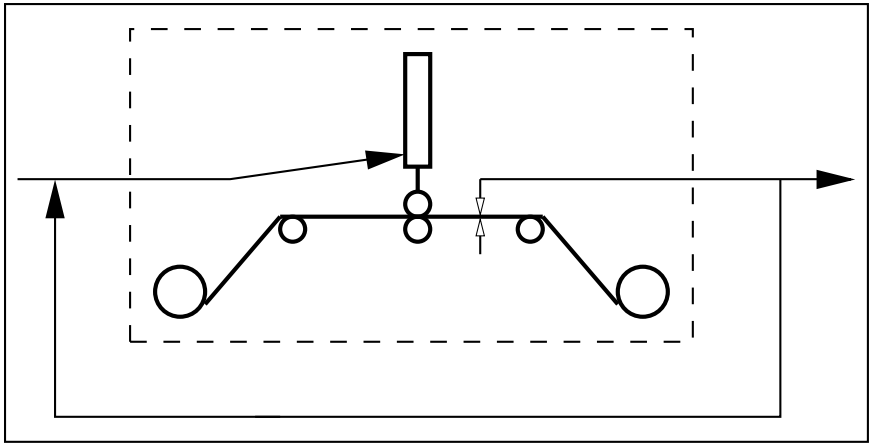
Conclusion

Operator control

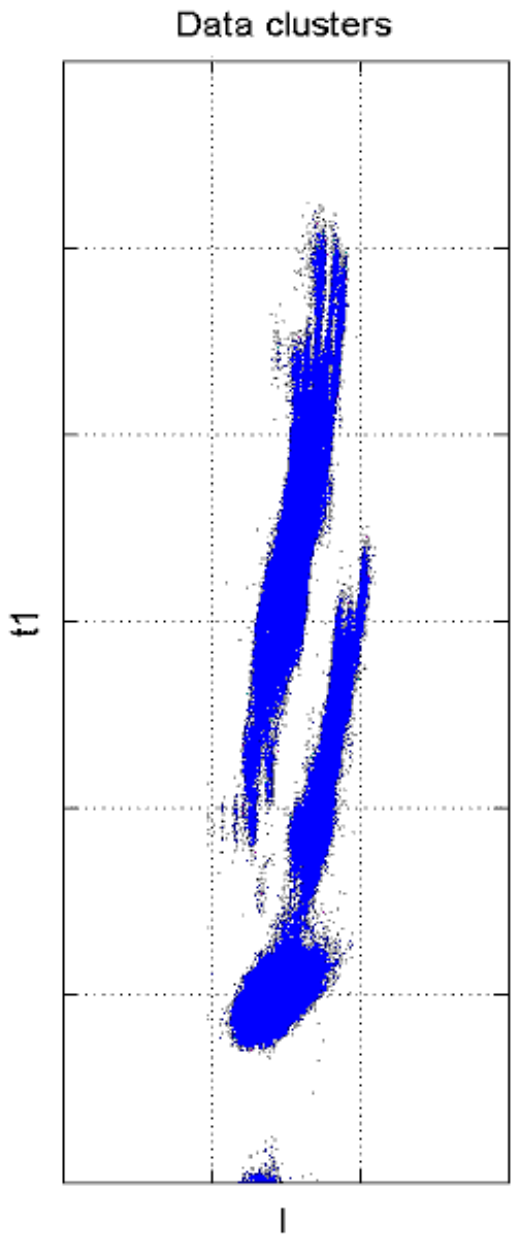


Ex-post analysis of recorded data reveals:

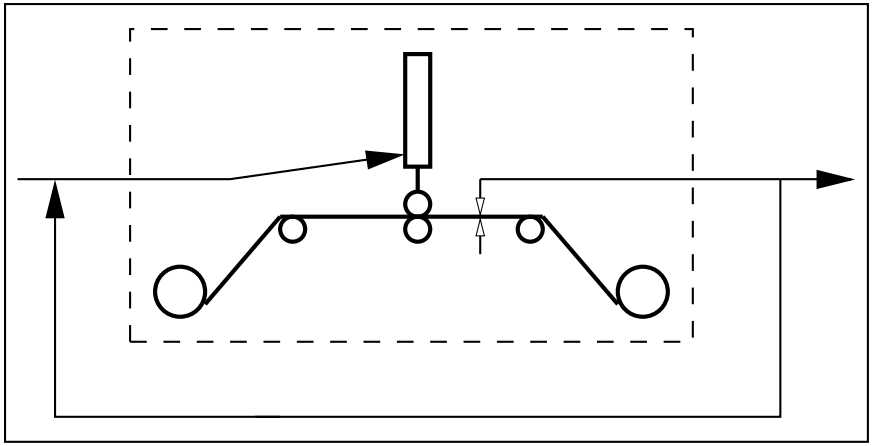
Operator control



Ex-post analysis of recorded data reveals: clusters?

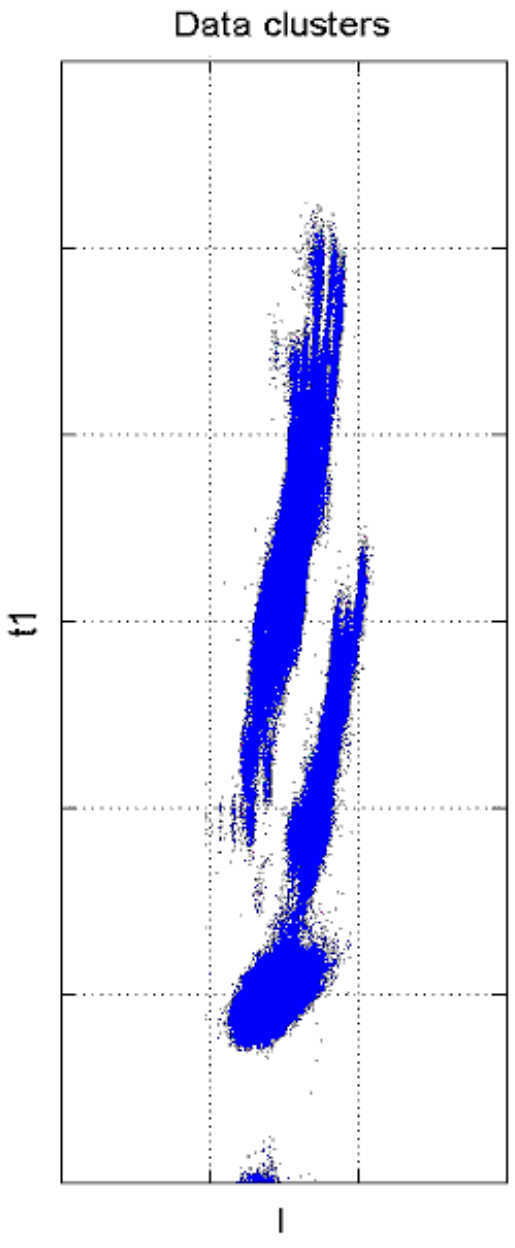


Operator control

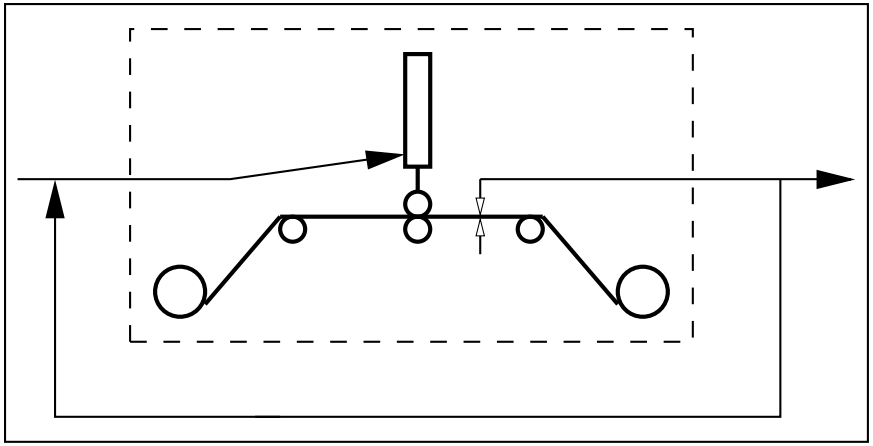


Ex-post analysis of recorded data reveals: clusters?

- working modes of the machine,
- people in the loop,
- operators are free to choose 'set points',
- distinct performance of different workers in terms of quality.



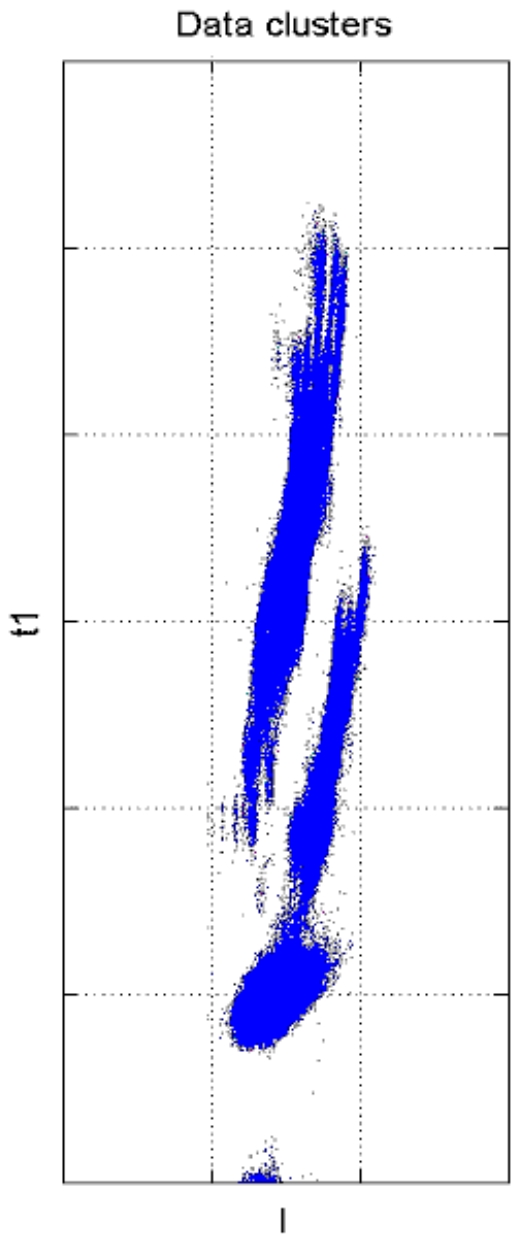
Operator control



Ex-post analysis of recorded data reveals: clusters?

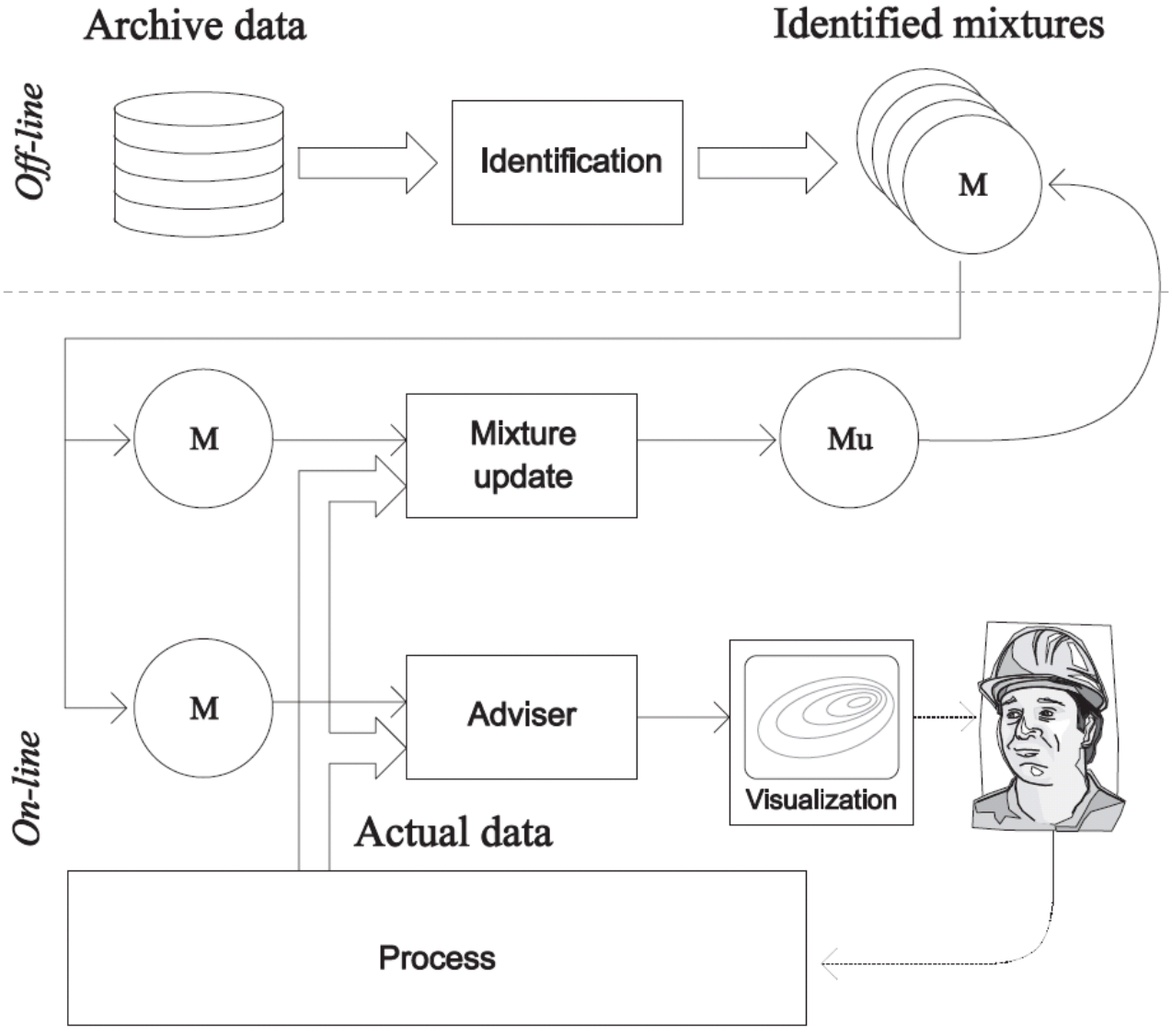
- working modes of the machine,
- people in the loop,
- operators are free to choose 'set points',
- distinct performance of different workers in terms of quality.

Basic idea: guide inexperienced operators to better set points.



EU project ProDaCTool, 2000–2002

U.of Reading, UTIA, TCD, Compureg, KOR



Results

Off-line tasks:

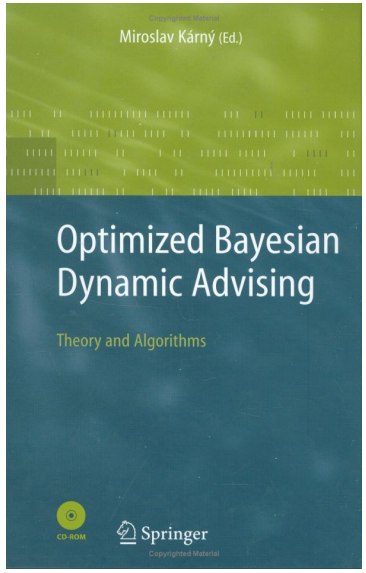
- Prior elicitation,
- Model structure selection (variable selection, covariance structure selection).

On-line:

- Recursive estimation of mixtures of ARX models,
- target elicitation,
- selection of visualization variables,
- recommendation generation.

Methodology:

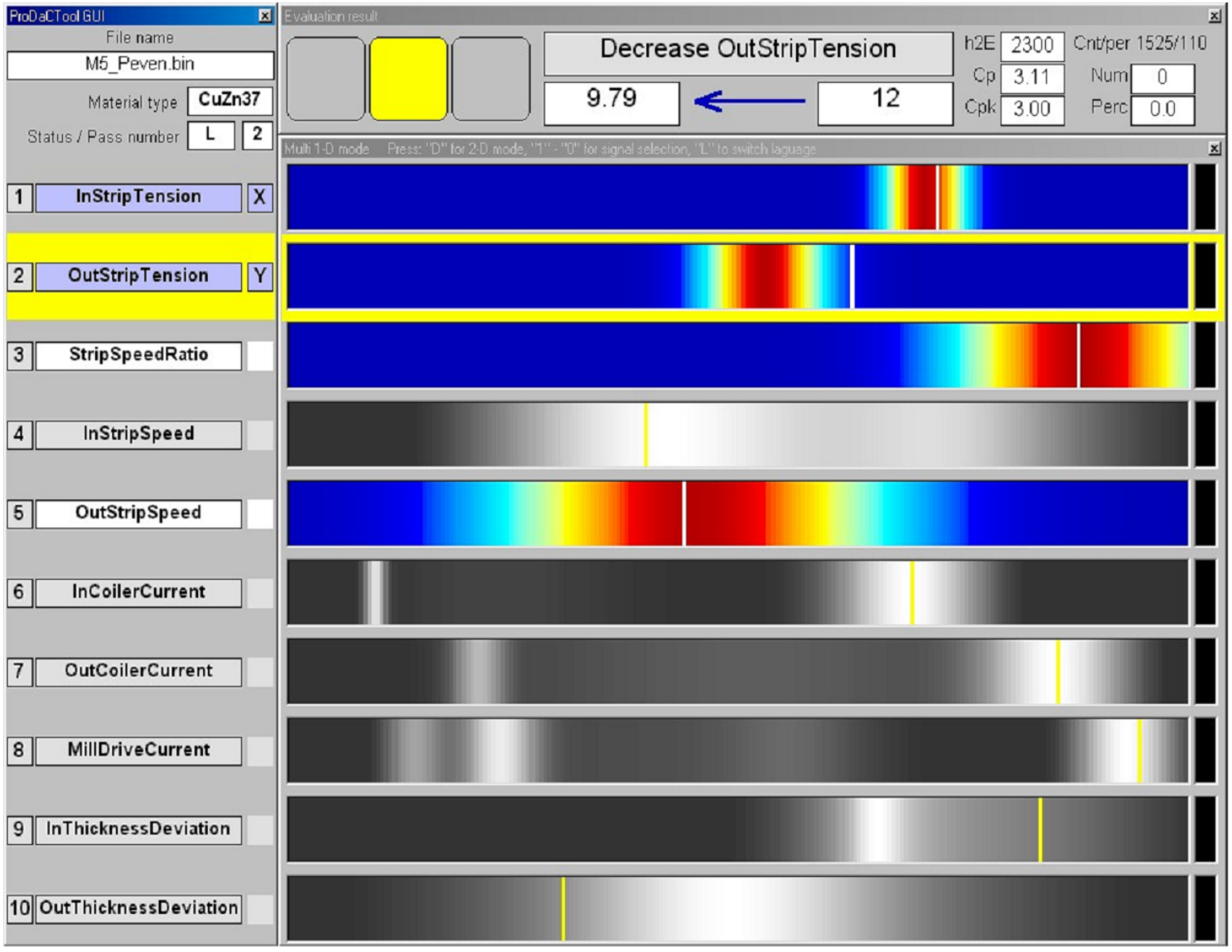
- ARX-based algorithms
- Kullback-Leibler divergence optimization.



M. Karny et. al, Springer 2006, 552pp.

Software: Mixtools, 1.8MB of C and Matlab code

Operator system GUI



Recursive estimation of a mixture model

EM algorithm: maximum likelihood

$$f(y_t|\Theta,\overline{\psi}_t) = \int f(y_t, l_t|\Theta,\overline{\psi}_t)dl_t = \sum_{i=1}^{c} f_i(y_t|\theta^{(i)},\psi_t^{(i)},l_t)\underbrace{f(l_t)}_{w_i}$$

'Missing data': component label, l_t,
Parameters: parameters of components, θ

E-step: compute expectation
$f(\theta|H) \propto \int f(l_t|y_{(1:t)}, u_{(1:t)}, \hat{\Theta}_t) \ln f(l_t, \Theta|H)dl_t$

M-step: find $\hat{\theta}_t = \arg\min_\theta f(\theta|H)$.

Interpretation as minimization of KL-divergence, Csiszar and Tusnady [1984], allows for recursive version.

Unsatisfactory results... after processing GB of data in Compureg and using in advising.

Quasi-Bayes estimation

Titterington et al. [1985], extended for ARX Kárný et al. [1998] using heuristic argument.

Recursive EM:

E-step: compute expectation
$$f(\theta|H) = \int f(l_t|y_{(1:t)}, u_{(1:t)}, \hat{\Theta}_{t-1}) \ln f(l_t, \Theta|H) dl_t$$

M-step: find $\hat{\theta}_t = \arg\min_\theta f(\theta|H)$.

QB:

Update: $f(\theta|H) = \int f(l_t|y_{(1:t)}, u_{(1:t)}) \ln f(l_t, \Theta|H) dl_t$.

(In LD decomposition, marginal is easy to compute...)

Most of the results achieved during the project are based on this method. Still not very satisfactory.

Variational Bayes

Ensemble learning, free entropy minimization, naive mean field, Variational EM.

- for $n \to \infty$, VB $\to$ EM.

Functional minimization of Kullback-Leibler under conditional independence assumption:

$$f(\Theta, l_t|H) \approx \tilde{f}(\Theta|H)\,\tilde{f}(l_t|H).$$

Optimum reaching algorithm:

E-step: compute expectation $\tilde{f}(\Theta|H) = \int \tilde{f}(l_t|H) \ln f(l_t, \Theta|H) dl_t$

E2-step: find $\tilde{f}(l_t|H) = \int \tilde{f}(\Theta|H) \ln f(l_t, \Theta|H) d\Theta$

Recursive version and proofs: Sato [2001].
Remarks:

- lower bound on marginal likelihood (no need for BIC),
- allows more complex model of weights.

Still not quite there...

Projection Bayes

Variational Bayes (and EM) optimize KL divergence

$$\tilde{f}(\Theta|H) = \arg\min_{f(\Theta)} \mathrm{D}(\tilde{f}(\Theta|H)||f(\Theta|H)).$$

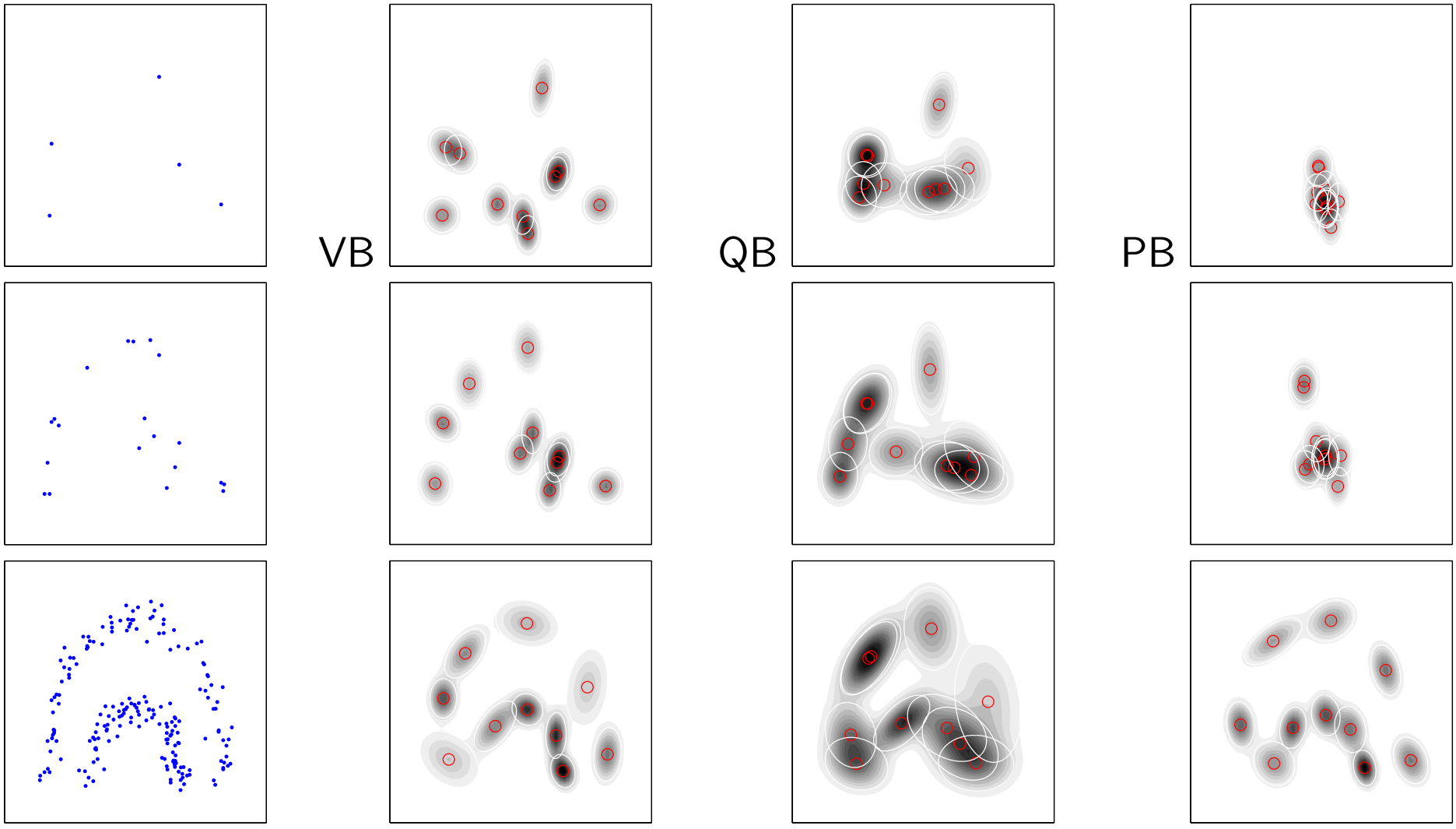
Better approximation should be obtained via

$$\tilde{f}(\Theta|H) = \arg\min_{f(\Theta)} \mathrm{D}(f(\Theta|H)||\tilde{f}(\Theta|H)).$$

Bernardo [1979], which does not suffer from local minima Amari et al. [2001].

In recursive mixture estimation it is closely related to moment matching (with a bit of numerical optimization), Andrýsek [2005].

Finally, we have got OK from Compureg.


Sim.
Prior
VB
QB
PB

What else is in the book...

- Prior positions of clusters: sequential runs. Posterior from one run is flattened and used as prior in the other.
- Number of components: branch and bound algorithm, (many variants)
  - branching is done according to clustering of residue (≈distance from centre) within each component,
  - bounding is triggered when KL divergence of two components falls below threshold.
- Each component estimates its covariance structure. Multivariate models are modelled as product of univariate factors

$$f(y_1, y_2, y_3) = f(y_1|y_2, y_3)f(y_2|y_3)f(y_3).$$

  Each univariate model in each components perform structure estimation (conditional independence). Uses technology of ARX models.
- Dimensionality reduction via common factors did not bring enough advantage.

Towards mixtures with dynamic weights

Control with mixtures with fixed weights

$$f(y_t|\Theta,\overline{\psi}_t) = \int f(y_t, l_t|\Theta,\overline{\psi}_t)dl_t = \sum_{i=1}^{c} f_i(y_t|\theta^{(i)},\psi_t^{(i)},l_t)\underbrace{f(l_t)}_{w_i}$$

fails because of unrealistic assumption of conditional independence

$$f(l_t|\Theta,\overline{\psi}_t) \equiv f(l_t).$$

Possible models:
Markov transition: $f(l_t|l_{t-1})$, estimation of transition matrix via VB, Šmídl and Quinn [2005].
Logistic regression: $f(l_t|\psi_t)$, can be estimated via numerical integration, Andrýsek [2005], or potentially by marginalized particle filter.

Challenge: design of control strategy for such models.

Conclusion

- Adaptive control is a challenging context for Bayesian techniques with many restrictions.
  - i.i.d. assumption does not hold,
  - restrictions such as zero time-delay requirement rules out very high percentage of statistical methods.
- Transfer of knowledge between statistics and control?
  - statistical methods needs to be 'robustified' for practise.
  - many engineering (works-for-me) solutions needs statistical justification,
- There are alternatives to EM algorithm. Some of them perform better especially in recursive algorithms.

Bibliography I

S. Amari, S. Ikeda, and H. Shimokawa. Information geometry of α-projection in mean field approximation. In M. Opper and D. Saad, editors, *Advanced Mean Field Methods*, Cambridge, Massachusetts, 2001. The MIT Press.

J. Andrýsek. Estimation of Dynamic Probabilistic Mixtures. Technical Report 2150, ÚTIA AV ČR, Praha, 2005.

J. M. Bernardo. Expected infromation as expected utility. *The Annals of Statistics*, 7(3):686–690, 1979.

A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

P. Ettler. An adaptive controller for škoda twenty-roll cold rolling mills. In *Proceedings of 2nd IFAC Workshop on Adaptive Systems in Control and Signal Processing*, pages 277–280, Lund, Sweden, 1986. Lund Institute of Technology.

A. H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, New York, 1979.

Bibliography II

M. Kárný, J. Kadlec, and E. L. Sutanto. Quasi-Bayes estimation applied to normal mixture. In J. Rojíček, M. Valečková, M. Kárný, and K. Warwick, editors, *Preprints of the 3rd European IEEE Workshop on Computer-Intensive Methods in Control and Data Processing*, pages 77–82, Prague, September 1998. ÚTIA AV ČR.

R. Kulhavý. Recursive nonlinear estimation: A geometric approach. *Automatica*, 26(3):545–555, 1990.

R. Kulhavý and M. B. Zarrop. On a general concept of forgetting. *International Journal of Control*, 58(4):905–924, 1993.

V. Peterka. Bayesian approach to system identification. In P. Eykhoff, editor, *Trends and Progress in System identification*, pages 239–304. Pergamon Press, Oxford, 1981.

M. Sato. Online model selection based on the variational Bayes. *Neural Computation*, 13:1649–1681, 2001.

D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical Analysis of Finite Mixtures*. John Wiley, New York, 1985.

V. Šmídl and A. Quinn. Mixture-based extension of the AR model and its recursive Bayesian identification. *IEEE Transactions on Signal Processing*, 53(9):3530–3542, 2005. URL `files/publ/tsp05.pdf`.