A Brief Comparison of Selected Forgetting Methods

Kamil Dedecius

10th International PhD Workshop on Systems and Control
Hluboká n./Vlt., Czech Republic

Institute of Information Theory and Automation

Department of Adaptive Systems

Academy of Sciences of the Czech Republic

Outline

- System Model
- Parameter Estimation
+ Estimation with Partial Forgetting
+ Estimation with Exponential Forgetting
+ Estimation with Alternative Forgetting
- Experiments
- Conclusions and Future Work

System Model

We suppose the system model

$$f(y_t|\psi_t, \theta), \quad t = 1, 2, \dots$$

y_t – model output
ψ_t – regression vector (inputs u_τ, outputs y_τ)
θ – vector of parameters (regr. coefficients)

or in a form of a **regression model**

$$y_t = \sum_{i=1}^{n} a_i y_{t-i} + \sum_{j=0}^{m} b_j u_{t-j} + c_t + e_t$$

$$m, n \in \mathbb{N}_0, \qquad a_i, b_j, c_t \in \theta, \qquad e_t \sim \mathcal{N}(0, r)$$

e.g. AR(1): $y_t = a y_{t-1} + c_t + e_t$

Parameter Estimation

Basic steps

- data update (incorporates new data)

$$f(\theta_t|d(t)) \propto f(y_t|\psi_t, \theta_t)\ f(\theta_t|d(t-1))$$

- time update (reflects $\theta_t \longrightarrow \theta_{t+1}$)

$$f(\theta_{t+1}|d(t)) = \int_{\theta^*} f(\theta_{t+1}|d(t), \theta_t)\ f(\theta_t|d(t))\, \mathrm{d}\theta_t$$

where $d_t = (u_t, y_t), \qquad d(t) = (d_1, \ldots, d_t)$

Parameter variability and time update:

- $\theta_{t+1} = \theta_t$ – 'formal' step
- $\theta_{t+1} \approx \theta_t$ – slowly varying parameters – we need forgetting

Exponential Forgetting

- AKA Time-weighted least squares (TWLS)
- AKA Flattening of the posterior pdf
  - by *forgetting factor* $\lambda \in (0, 1]$
  - usually $\lambda \geq 0.95$
- In general form:

$$f(\theta_{t+1}|d(t)) = [f(\theta_t|d(t))]^\lambda$$

- In Gaussian model:

$$V_t = \lambda V_{t-1}$$
$$\nu_t = \lambda \nu_{t-1}$$

Alternative Forgetting

- AKA Stabilized exponential forgetting (SEF)
  - *forgetting factor* $\lambda \in [0, 1]$
  - two pdfs f_1 and f_2 for θ
- In general form:

$$f(\theta_{t+1}|d(t)) \propto [f_1(\theta|d(t))]^{\lambda}[f_2(\theta|d(t))]^{1-\lambda}$$

$$\min_f [\lambda \mathrm{D}\left(f||f_1\right) + (1-\lambda)\mathrm{D}\left(f||f_2\right)]$$

- In Gaussian model:

$$V_t = \lambda V_{t-1} + (1-\lambda)V_A$$
$$\nu_t = \lambda \nu_{t-1} + (1-\lambda)\nu_A$$

Partial Forgetting (PFM)

The principle

- The parameters have some true distribution with pdf ${}^{T}f$
  - which is unknown
  - but we can make hypotheses about it
  - $\longrightarrow$ and use them for approximation

Hypotheses

- No parameter varies – the filtered pdf

$$H_0: \; \mathrm{E}\left[{}^{T}f(\theta|d(t))|\theta, d(t), H_0 \right] = f(\theta|d(t))$$

- All parameters vary – an alternative pdf

$$H_1: \; \mathrm{E}\left[{}^{T}f(\theta|d(t))|\theta, d(t), H_1 \right] = f_A(\theta)$$

Partial Forgetting (PFM) – cont.

- A subset of parameters vary
  - $\theta_\alpha \in \theta$ – params. that do not vary
  - $\theta_\beta = \theta \setminus \theta_\alpha$ – params. that vary
  - ...and use the *chain rule* (*)

$$H_j: \; \mathrm{E}\left[{}^{T}f(\theta|d(t))|\theta, d(t), H_j \right] = f(\theta_\alpha|\theta_\beta, d(t)) f_A(\theta_\beta)$$

  - Theoretically up to 2^n hypotheses.
- Each hypothesis has assigned a weight (probability)

$$\lambda_j \in [0,1]; \quad \sum_j \lambda_j = 1, \quad j = 0, 1, \ldots$$

================================

$$f(\theta) = f(\theta_1, \ldots, \theta_n) = f(\theta_1) \prod_{i=2}^{n} f(\theta_i|\theta_{i-1}, \ldots, \theta_1) \qquad (*)$$

Approximation

A true pdf ${}^{T}f$ (or its expectation) can be expressed as a convex combination of the hypothetic densities:

$$\sum_j \lambda_j \mathrm{E}\left[\, {}^{T}f(\theta|d(t))|\theta, d(t), H_j\right]$$

…and then approximated by $\tilde{f}$

$$\mathrm{D}\left({}^{T}f(\theta) \middle|\middle| \tilde{f}(\theta) \right) = \int {}^{T}f(\theta) \ln \frac{{}^{T}f(\theta)}{\tilde{f}(\theta)} \mathrm{d}\theta$$

As we don't know ${}^{T}f$, we use the mixture and search for $\tilde{f}$.

Comparisons

The three methods were compared

- AR(1) model for simulated data

$$y_{t+1} = \theta_1 + \theta_2 y_t$$

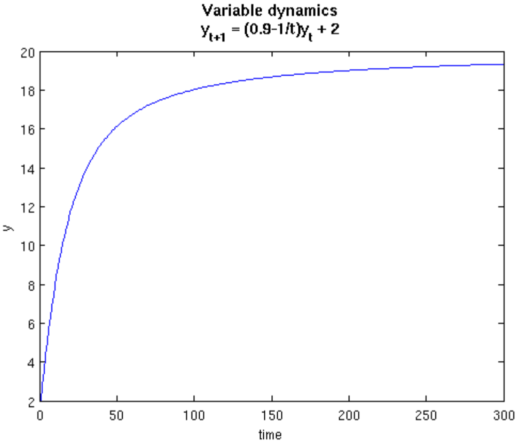
- the best weights/factors were searched
- alternative pdf $\longrightarrow$ flat prior
- criterion: Relative prediction error

$$RPE = \frac{1}{s}\sqrt{\frac{\sum_{i=1}^{T}(y_{p;i} - y_i)^2}{T}}$$

where y_i denotes the real system output, $y_{p;i}$ is the predicted output and s is the sample standard deviation of data on horizon T.

Time-varying dynamics

$$y_{t+1} = (0.9 - 1/t)y_t + 2, \quad t = 1, 2, \ldots, 300$$



Time-varying dynamics

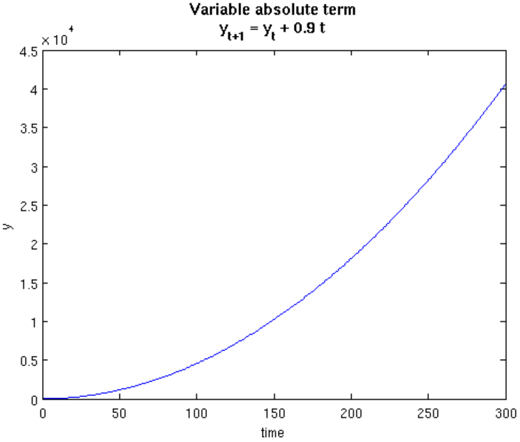
$$y_{t+1} = (0.9 - 1/t)y_t + 2, \quad t = 1, 2, \ldots, 300$$

Table: Time-varying dynamics: one step-ahead prediction of time series.

Method	Weight(s)	RPE
Exponential	0.95	0.00336
Alternative	0.4078	0.00085
Partial	[0.2443, 0.1435, 0.6122, 0]	0.00061

Time-varying absolute term

$$y_{t+1} = y_t + 0.9t, \quad t = 1, 2, \dots, 300$$



Time-varying absolute term

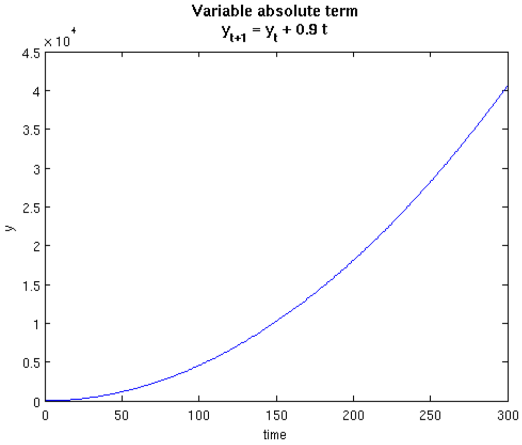
$$y_{t+1} = y_t + 0.9t, \quad t = 1, 2, \ldots, 300$$

Table: Time-varying absolute term: one step-ahead prediction of time series.

Method	**Weight(s)**	**RPE**
Exponential	0.95	19.564e-05
Alternative	0.001	7.0712e-05
Partial	[0.2941,0.0086, 0.6973]	6.436e-05

Time-varying absolute term and dynamics

$$y_{t+1} = (1 + 10^{-4}t)y_t + 10^{-3}t, \quad t = 1, 2, \ldots, 300$$



Time-varying absolute term and dynamics

$$y_{t+1} = (1 + 10^{-4}t)y_t + 10^{-3}t, \quad t = 1, 2, \ldots, 300$$

Table: Time-varying both parameters: one step-ahead prediction of time series.

Method	Weight(s)	RPE
Exponential	0.95	33.478e-05
Alternative	0.001	9.789e-05
Partial	[0.731,0.0020,0.2490,0]	9.216e-05

Conclusions and Future Work

Conclusions

+ The PFM method leads to the best results.
+ The AF method was very succesfull too.
+ The most basic EF method led to worse results.

However. . .

+ The EF is very simple!
- The PFM is very complicated in comparison to the others.
+ However, PFM can fully elliminate the blow-up phenomenon, when the covariance grows w/o bounds.

Future work

- Method for online optimization of hypotheses' weights of PFM
- Method for constructing appropriate alternative pdfs for PFM

The End

Thank you for your attention