Estimation with Partial Forgetting

Lenka Pavelková, Kamil Dedecius, Ivan Nagy

ÚTIA AV ČR, Department of Adaptive Systems

Model and parameter pdf

Normal regression model pdf

$$m\left(y_t|\psi_t,\Theta\right) \propto r^{-0.5}\exp\left\{-\frac{1}{2r}\left(y_t-\theta'\psi_t\right)^2\right\} =$$

$$= r^{-0.5}\exp\left\{-\frac{1}{2r}\begin{bmatrix}-1\\ \theta\end{bmatrix}'\begin{bmatrix}y_t\\ \psi_t\end{bmatrix}\begin{bmatrix}y_t\\ \psi_t\end{bmatrix}'\begin{bmatrix}-1\\ \theta\end{bmatrix}\right\}$$

Self-reproducing parameter pdf

$$p\left(\Theta|d\left(t\right)\right) \propto r^{-\nu_t}\exp\left\{-\frac{1}{2r}\begin{bmatrix}-1\\ \theta\end{bmatrix}'V_t\begin{bmatrix}-1\\ \theta\end{bmatrix}\right\}$$

Estimation with pdfs

Bayes rule

$$p_{t-1|t} \propto m_t \, p_{t-1|t-1} \quad - \quad \text{filtration}$$

$$p_{t|t} = p^{\lambda}_{t-1|t} \quad - \quad \text{prediction}$$

Where, prediction means forgetting.

Estimation with statistics

Estimation

$$V_t = V_{t-1} + \begin{bmatrix} y_t \\ \psi_t \end{bmatrix} \begin{bmatrix} y_t \\ \psi_t \end{bmatrix}'$$

Estimation with forgetting

$$V_{t-1|t} = V_{t-1|t-1} + \begin{bmatrix} y_t \\ \psi_t \end{bmatrix} \begin{bmatrix} y_t \\ \psi_t \end{bmatrix}', \qquad V_{t|t} = \lambda V_{t-1|t}$$

Estimation with factorized information matrix

$$L_t' D_{t-1|t} L_t = L_{t-1}' D_{t-1|t-1} L_{t-1} + \begin{bmatrix} y_t \\ \psi_t \end{bmatrix} \begin{bmatrix} y_t \\ \psi_t \end{bmatrix}', \quad D_{t|t} = \lambda D_{t-1|t}$$

Parameter pdf decomposition

Exponent of parameter pdf

$$\begin{bmatrix} -1 \\ \theta \end{bmatrix}' L_t' D_{t|t-1} L_t \begin{bmatrix} -1 \\ \theta \end{bmatrix} = \left(L_t \begin{bmatrix} -1 \\ \theta \end{bmatrix} \right)' D_{t|t-1} \left(L_t \begin{bmatrix} -1 \\ \theta \end{bmatrix} \right) =$$

$$= \left(\begin{bmatrix} -1 \\ \tilde{\theta} \end{bmatrix} \right)' D_{t|t-1} \left(\begin{bmatrix} -1 \\ \tilde{\theta} \end{bmatrix} \right) = D_{1;t|t-1} + \sum_{i=2}^{n} D_{i;t|t-1} \tilde{\theta}_i^2$$

Filtered parameter

$$\begin{bmatrix} -1 \\ \tilde{\theta} \end{bmatrix} = L_t \begin{bmatrix} -1 \\ \theta \end{bmatrix} \quad \rightarrow \quad \tilde{\theta}_k = \sum_{i=2}^{k} L_{k,i} \theta_i$$

Decomposition to conditional pdfs

Back to forgetting

Prediction step - statistics

$$D^{\lambda_1}_{1;t|t-1} + \sum_{i=2}^{n} D^{\lambda_i}_{i;t|t-1}\tilde{\theta}^2_i = \begin{bmatrix} -1 \\ \theta \end{bmatrix}' L'_t D_t L_t \begin{bmatrix} -1 \\ \theta \end{bmatrix}$$

where

$$D_t = \text{diag}\left(\lambda_1 D_{1;t|t-1},\ \lambda_2 D_{2;t|t-1}, \cdots,\ \lambda_n D_{n;t|t-1},\right)$$

What is forgotten, are the conditional parameter pdfs

Theoretical justification

A vector of parameters $\Theta = [\Theta_1, \Theta_2, \cdots, \Theta_n]$ is given. Its items either vary in time or they are constant. We are to construct its description

$$f(\Theta) = f(\Theta_1, \Theta_2, \cdots, \Theta_n) = f(\Theta_1|\Theta_{2:n})\, f(\Theta_2|\Theta_{3:n}) \cdots f(\Theta_n)$$

on condition, that we know the following $N = 2^n$ hypotheses

H_1 - all parameters Θ_i do not vary $\cdots$ with probability α_1

H_2 - Θ_1 varies, others do not $\cdots$ with probability α_2

. . . . . .

H_N - all parameters vary $\cdots$ with probability α_N

Theoretical justification

If the parameter Θ_i varies, it is described by $f_U(\Theta_i|\Theta_{i+1:n})$. If it is constant, its description is given by $f_S(\Theta_i|\Theta_{i+1:n})$. Thus, for the j-th hypothesis, the pdf $f_j(\Theta)$ is a product of a corresponding combination of f_U and f_S.

The resulting pdf $\hat{f}(\Theta)$ is such one, that minimizes the KL distance

$$\min_{\hat{f}} E\left[D\left(f,\hat{f}\right)\right] = \min_{\hat{f}} E\left[\int f(\Theta)\log\frac{f(\Theta)}{\hat{f}(\Theta)}d\Theta\right]$$

Solution

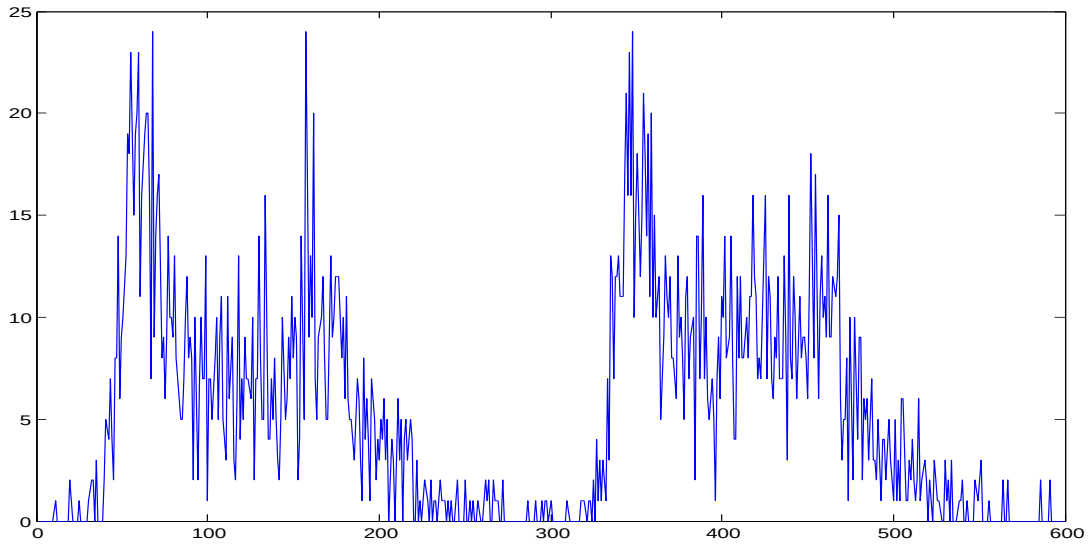
$$\hat{f} = \prod_{i=1}^{N} f_i^{\alpha_i}$$

Example

Consider first order regression model

$$y_t = ay_{t-1} + k + e_t$$

and transportation data - intensities of the traffic flow



Example

We suppose the model should work so that

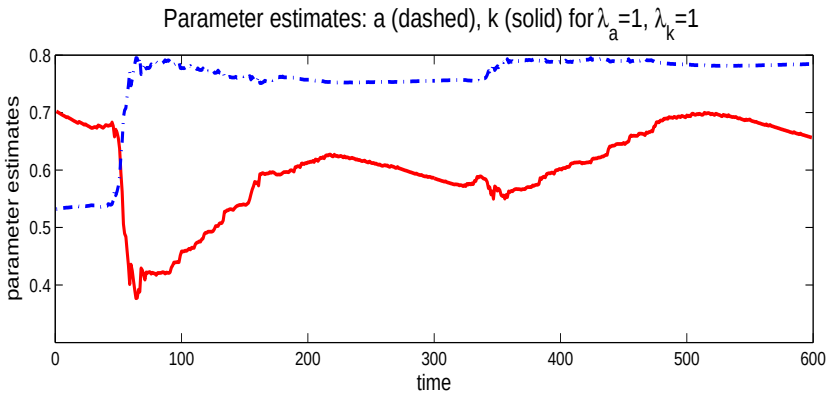
- the absolute term k changes due to the changing data
- the other parameters a and noise variance r stay constant

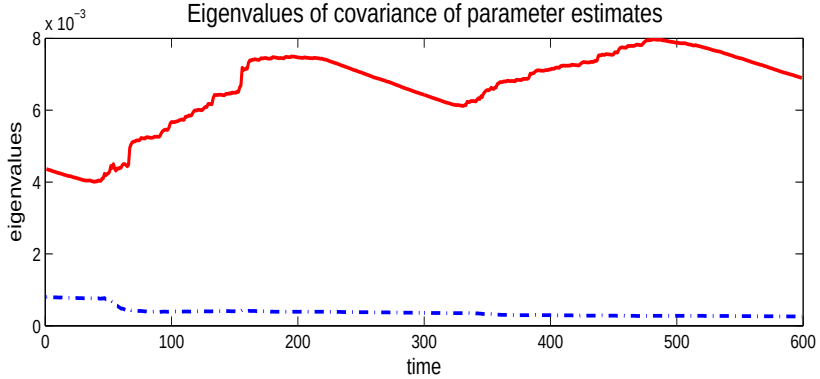
$\Longrightarrow$ two different forgetting factors

- λ_U - "small" belonging to absolute term
- λ_S - "big" for the rest of the parameters.

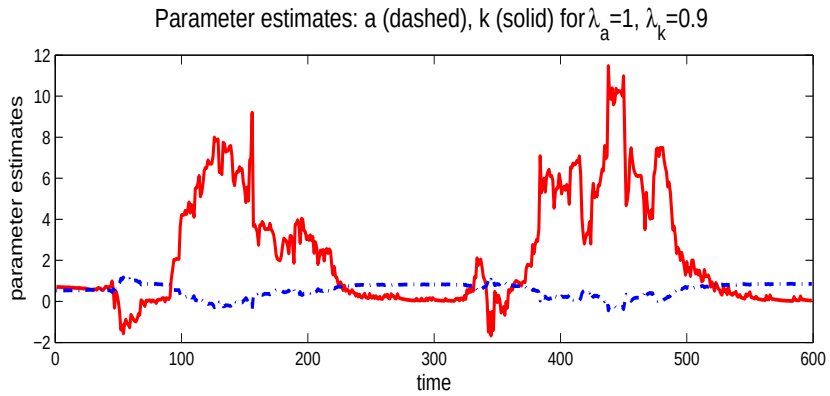
Example

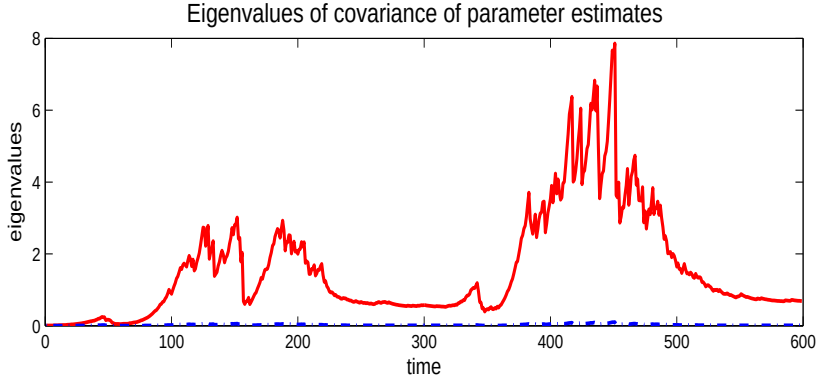
$\lambda_U = \lambda_S = 1$





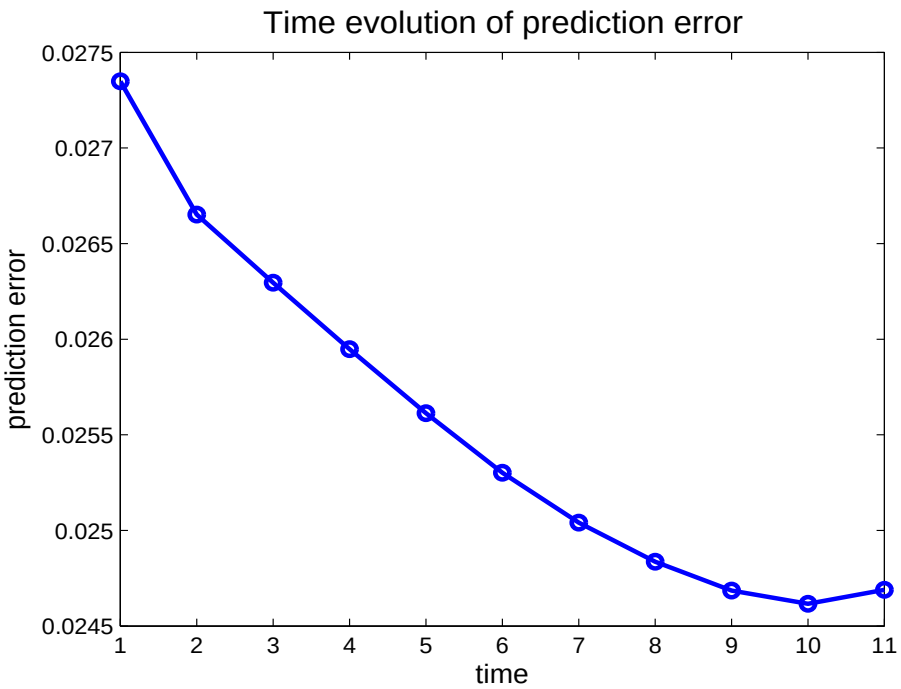
$\lambda_U = 0.9,\ \lambda_S = 1$

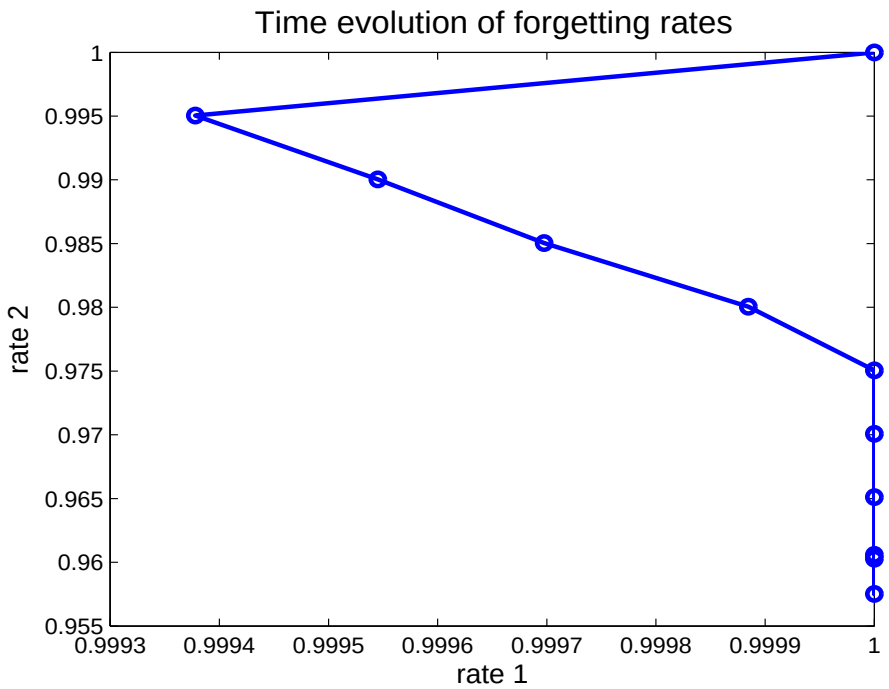




Example

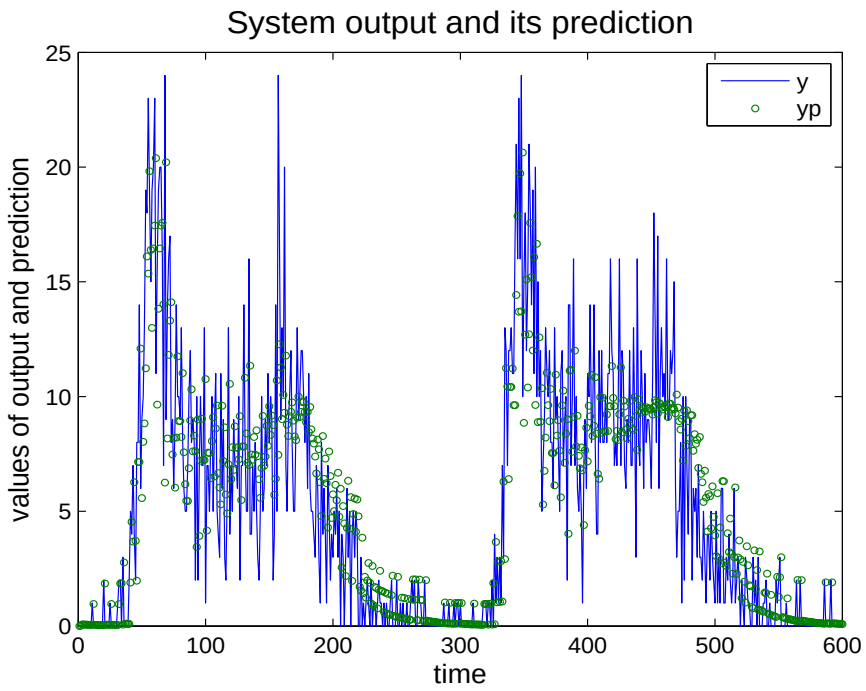
Search for optimal forgetting factors

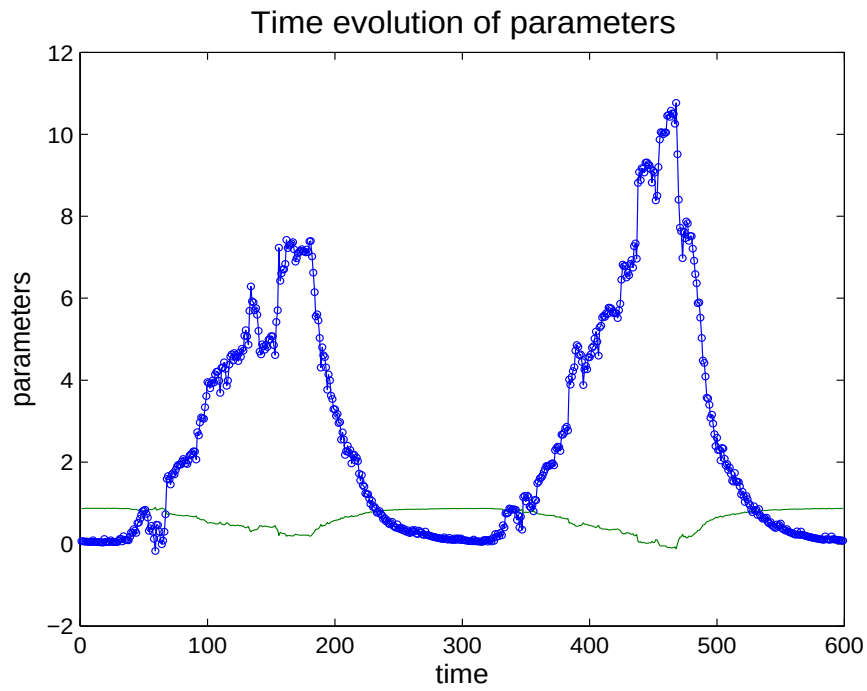




Example

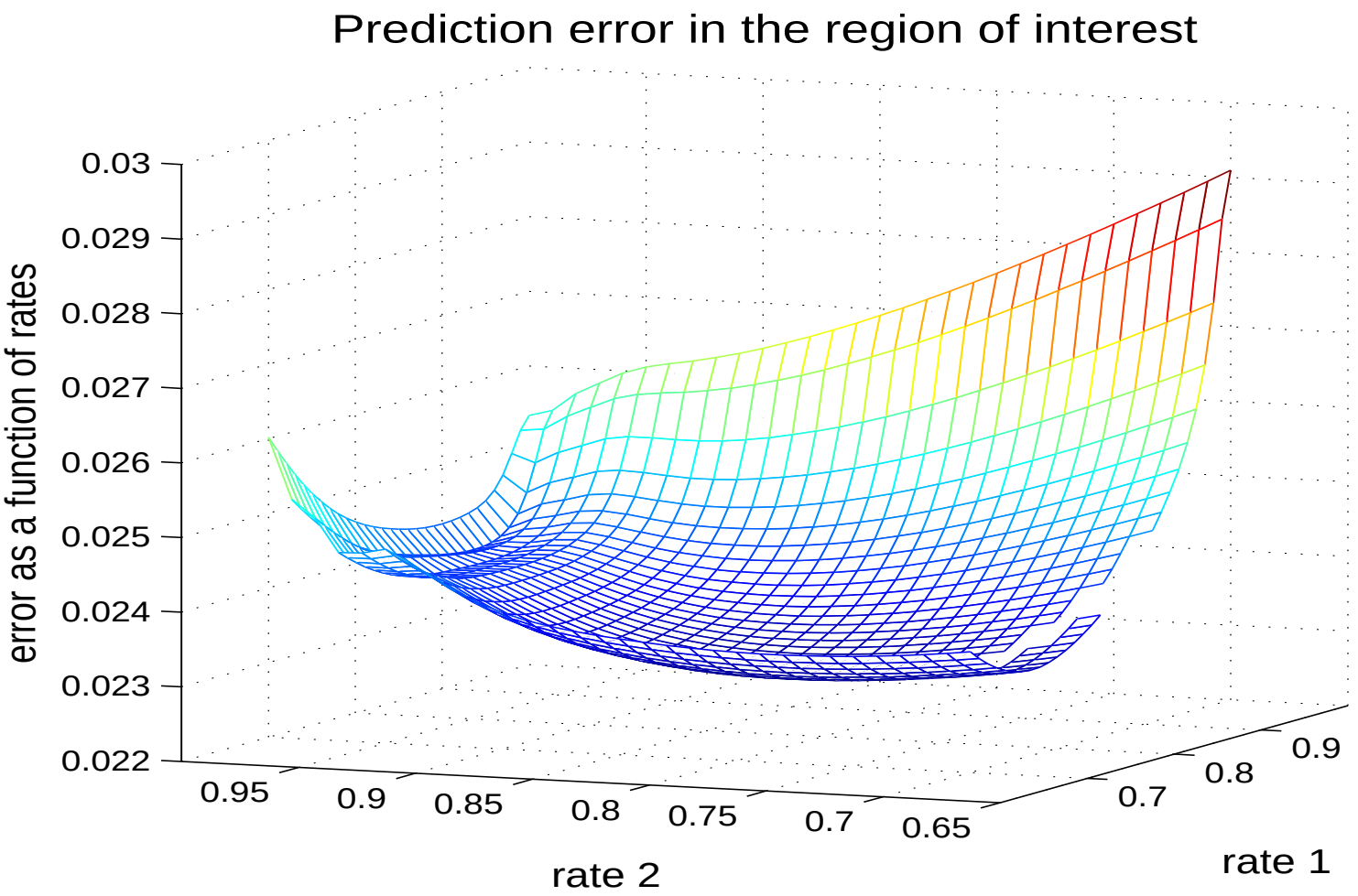
Results with optimal forgetting factors





Example

Criterion for optimal search - normalized variance of prediction error



Conclusions

- Attempt, based on conditional pdfs and normal KL distance
- Other approaches are being followed
  - for marginal pdfs
  - for the reverse KL distance