Data dropouts in Bayesian model averaging for application in cold rolling

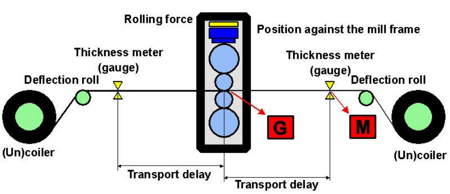
(experiments with potential methods)

Ladislav Jirsa

ÚTIA AV ČR, v.v.i.
Prague, Czech Republic

6. 12. 2010, ÚTIA, seminar AS

Cold metal rolling



Problems and tasks

To control the rolling plant, the controlled quantity (output thickness of the metal strip) must be known.

Problems

- metal thickness cannot be measured in the rolling gap
- output metal thickness is measured with a high transport delay (19, 123)
- faulty and noisy data

Tasks

- predict the non-delayed output thickness (“soft sensor”)
- use mixture of multiple models with dynamic weights

The idea of BMA

We use K models $M^1, M^2, ..., M^K$ for prediction of the target quantity p

$$\text{data} \rightarrow \left\{ \begin{array}{lllll} \text{model } M^1 & \rightarrow & \text{predictor } p|M^1 & \rightarrow & \text{weight } \omega^1 \\ \text{model } M^2 & \rightarrow & \text{predictor } p|M^2 & \rightarrow & \text{weight } \omega^2 \\ \cdots & & \cdots & & \cdots \\ \text{model } M^K & \rightarrow & \text{predictor } p|M^K & \rightarrow & \text{weight } \omega^K \end{array} \right\} \rightarrow p = \sum_{i=1}^{K} \omega^i \, p|M^i$$

Dynamic model averaging — allows time changes of parameters and weights

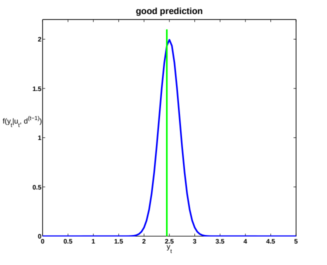
Estimation and prediction

data $d_t = [y_t, u_t']'$ parameter Θ_t

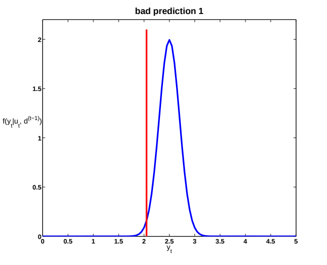
- single model M
  - estimation of slowly varying parameters
    - time update: forgetting (exponential, stabilized, *partial*, not linear), parameter λ
    - data update: Bayes formula
  - prediction: $f\left(y_t|u_t, d^{(t-1)}, M\right) = \int f\left(y_t|u_t, d^{(t-1)}, \Theta_t, M\right)\; f\left(\Theta_t|d^{(t)}, M\right)\, \mathrm{d}\Theta_t$

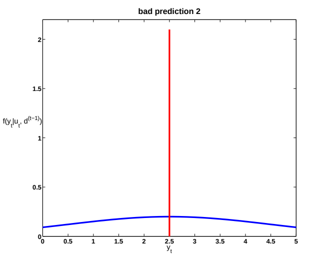
- multiple models M^i, $i = 1, ..., K$
  - estimation and prediction for each model
  - estimation of varying weights $\omega^i_{t|t}$ based on predictions
    - time update: linear forgetting, $\omega^i_{t-1|t-1} \longrightarrow \omega^i_{t|t-1}$, parameter α
    - data update: Bayes formula $\omega^i_{t|t} \propto f(y_t|u_t, d^{(t-1)}, M^i)\, \omega^i_{t|t-1}$

Quality of prediction



- measured data give high value of predictive density → "good" prediction
- measured data give low value of predictive density → "bad" prediction





Linear regression model with normal noise

- model $y_t = \vartheta_t' \psi_t + e_t$ noise $e_t \sim \mathcal{N}(0, r_t)$ parameter $\Theta_t = [\vartheta_t', r_t]'$
- regressor $\psi_t \subset \{u_t, y_{t-1}, u_{t-1}, \ldots, y_{t-\partial}, u_{t-\partial}\}$ data vector $\phi_t = [y_t, \psi_t']'$
- expressing belief in data: data weights $w_t \in \langle 0, 1 \rangle$
- conjugate system — Gauss-inverse-Wishart

$$\left.\begin{aligned} f\left(\Theta_t | d^{(t)}\right) &\equiv f\left(\Theta_t | V_t, \nu_t\right) \\ V_t &= \lambda V_{t-1} + w_t^2\ \phi_t \phi_t' \\ \nu_t &= \lambda \nu_{t-1} + w_t \end{aligned}\right\} \text{exponential forgetting with data weights}$$

- decomposition of extended information matrix $V_t = L_t' \, D_t \, L_t$
  - L_t — lower triangular with unit diagonal
  - D_t — diagonal matrix
- estimate of noise variance $\hat{r}_t = \frac{D_{11,t}}{\nu_t - 2}$ predictive variance $\hat{r}_{p,t+1|t} > \hat{r}_t$

Weights depend on predictive *pdf*, its variance depends on D_{11}.

Dealing with data dropouts

Let be available quantity $a_{jt} \in \langle 0, 1 \rangle$ expressing reliability of j-th data channel in time t
For ϕ_t^i get a_t^i as a product of corresponding entries of a_{jt}

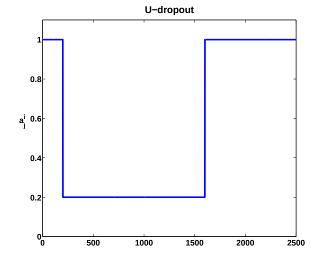
These methods were tested as possible tools:

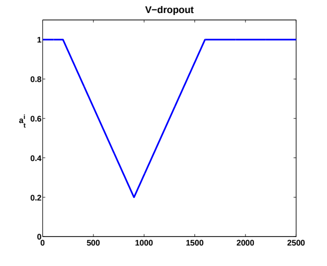
1. assigning zero values to the data
2. direct modification of $D_{11,t}$
3. modification of forgetting factor λ
4. assigning weights to the data vector ϕ_t

Each method affects a different link of the chain:

$$\text{data} \rightarrow \left\{ \begin{array}{lclcl} \text{model } M^1 & \rightarrow & \text{predictor } p|M^1 & \rightarrow & \text{weight } \omega^1 \\ \text{model } M^2 & \rightarrow & \text{predictor } p|M^2 & \rightarrow & \text{weight } \omega^2 \\ \dots & & \dots & & \dots \\ \text{model } M^K & \rightarrow & \text{predictor } p|M^K & \rightarrow & \text{weight } \omega^K \end{array} \right\} \rightarrow p = \sum_{i=1}^{K} \omega^i \, p|M^i$$

Types of data dropouts used in the experiments





- u-dropout, v-dropout
- $b_1 \leq a_t^i \leq b_2$
- both b_1 and b_2 can be chosen
- in the experiments, $b_2 = 1$
- here, $b_1 = 0.2$

Models used for mixing

These models were used for experiments:

1. derived from mass-flow principle, $\psi = [v_r, h_1 v_r, 1]'$
2. derived from gaugemeter principle — linear force, $\psi = [z, F, 1]'$
3. simplest “gray box” model, $\psi = [h_1, z, 1]'$
4. another “gray box” model, $\psi = [h_1, z, v_r, 1]'$
5. derived from gaugemeter principle — quadratic force, $\psi = [z, F, F^2, 1]'$

Assigning zero values to the data

Principle:

- initial attempt for binary $a_{jt} \in \{0, 1\}$
- it leaves the predictive variance to its fate
- based on assumption that zero-valued data channels will decrease prediction ability and hence the weights of the affected model

Advantages:

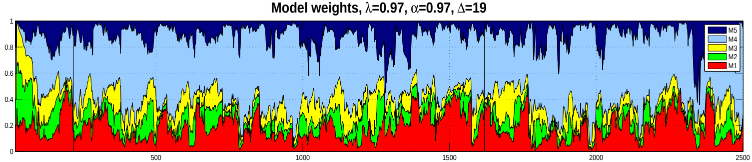
- natural and implicit mechanism

Disadvantages:

- it does not work

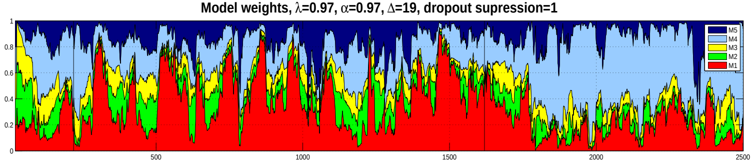
Experiments with the method

No dropout



mixture is better than the single model:	0.47981	0.59736	0.54898	0.46541	0.57177	0
maximal absolute prediction error:	33.9325	13.7322	42.2052	42.1793	17.1499	22.9287
mean prediction error:	−0.038359	0.014908	−0.02256	−0.052713	0.012046	−0.040253
median of prediction error:	0.010877	0.10452	0.023774	−0.044317	0.092584	−0.034884
std. dev. of prediction error:	3.2356	3.5536	3.5977	3.3209	3.5597	3.2919
mean weights of the model:	0.18562	0.096605	0.11604	0.47951	0.11463	

Dropout in h_1 between 200 and 1600 (zeroes), inputs of models 3 and 4 affected



mixture is better than the single model:	0.4926	0.60896	0.54938	0.4926	0.58936	0
maximal absolute prediction error:	33.9325	13.7322	42.2052	42.1793	17.1499	22.9287
mean prediction error:	−0.038359	0.014908	0.29614	0.35426	0.012046	0.045001
median of prediction error:	0.010877	0.10452	0.29818	0.40116	0.092584	0.10583
std. dev. of prediction error:	3.2356	3.5536	3.5462	3.3187	3.5597	3.2557
mean weights of the model:	0.32282	0.11469	0.10014	0.31299	0.14175	

Direct modification of D_{11}

Principle:

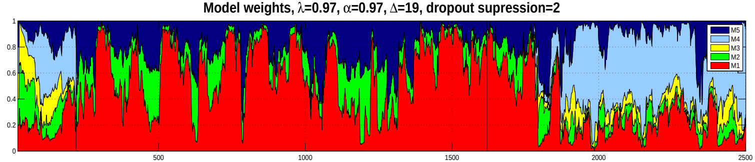
- model weight $\omega^i_{t|t}$ is affected by the predictive variance given by $D^i_{11,t}$
- significant increase of the predictive variance
- $D_{11,t} \leftarrow D^i_{11,t} + (1 - a^i_t)\ C,$
- C is chosen to be "much" greater than original $D^i_{11,t}$ to increase predictive variance

Effect:

- immediate suppression of the respective $\omega^i_{t|t}$ if $a^i_t < 1$
- if $a^i_t = 1$, return to the regular operating state is slow and it is given by λ and C

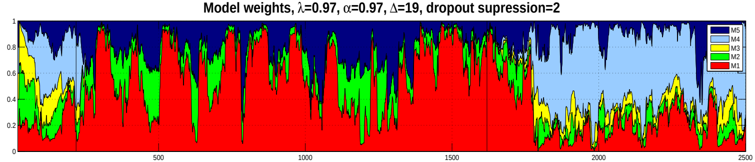
Experiments with the method I

U-dropout in h_1 between 200 and 1600, $(b_1, b_2) = (0, 1)$, $C = 10^3$



mixture is better than the single model:	0.47421	0.59456	0.53699	0.45742	0.58457	0
maximal absolute prediction error:	33.9325	13.7322	42.2052	42.1793	17.1499	22.9287
mean prediction error:	−0.038359	0.014908	−0.02256	−0.052713	0.012046	−0.015552
median of prediction error:	0.010877	0.10452	0.023774	−0.044317	0.092584	−0.018707
std. dev. of prediction error:	3.2356	3.5536	3.5977	3.3209	3.5597	3.3064
mean weights of the model:	0.45886	0.14526	0.041522	0.16172	0.18505	

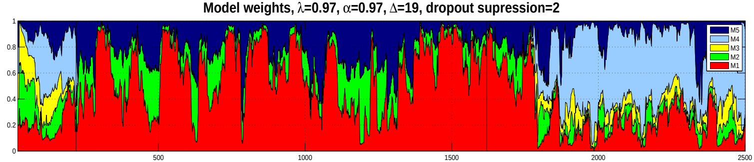
V-dropout in h_1 between 200 and 1600, $(b_1, b_2) = (0, 1)$, $C = 10^3$



mixture is better than the single model:	0.47181	0.59536	0.54018	0.45942	0.58537	0
maximal absolute prediction error:	33.9325	13.7322	42.2052	42.1793	17.1499	22.9287
mean prediction error:	−0.038359	0.014908	−0.02256	−0.052713	0.012046	−0.0058387
median of prediction error:	0.010877	0.10452	0.023774	−0.044317	0.092584	−0.015439
std. dev. of prediction error:	3.2356	3.5536	3.5977	3.3209	3.5597	3.3018
mean weights of the model:	0.44196	0.14138	0.047045	0.18662	0.1754	

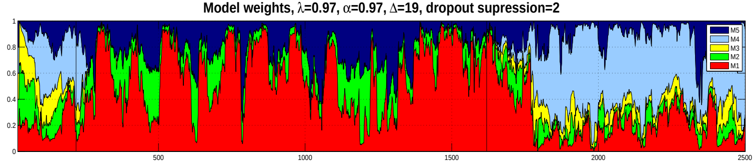
Experiments with the method II

U-dropout in h_1 between 200 and 600, $(b_1, b_2) = (0.5, 1)$, $C = 10^3$



mixture is better than the single model:	0.47421	0.59496	0.53898	0.45702	0.58457	0
maximal absolute prediction error:	33.9325	13.7322	42.2052	42.1793	17.1499	22.9287
mean prediction error:	−0.038359	0.014908	−0.02256	−0.052713	0.012046	−0.01249
median of prediction error:	0.010877	0.10452	0.023774	−0.044317	0.092584	−0.01871
std. dev. of prediction error:	3.2356	3.5536	3.5977	3.3209	3.5597	3.3038
mean weights of the model:	0.45382	0.14418	0.042959	0.17014	0.18131	

V-dropout in h_1 between 200 and 600, $(b_1, b_2) = (0.5, 1)$, $C = 10^3$

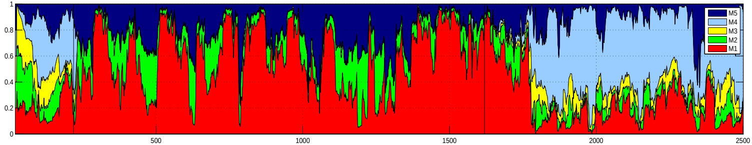


mixture is better than the single model:	0.47141	0.59696	0.53978	0.45982	0.58457	0
maximal absolute prediction error:	33.9325	13.7322	42.2052	42.1793	17.1499	22.9287
mean prediction error:	−0.038359	0.014908	−0.02256	−0.052713	0.012046	−0.0049387
median of prediction error:	0.010877	0.10452	0.023774	−0.044317	0.092584	−0.017578
std. dev. of prediction error:	3.2356	3.5536	3.5977	3.3209	3.5597	3.3021
mean weights of the model:	0.43723	0.14054	0.049018	0.19137	0.17425	

Experiments with the method III

U-dropout in h_1 between 200 and 1600, $(b_1, b_2) = (0.5, 1)$, $C = 10^2$

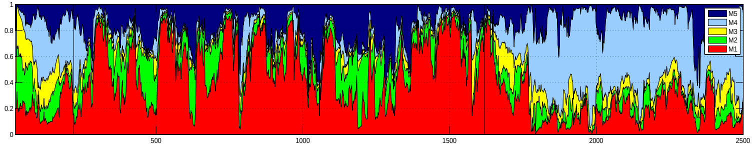
Model weights, λ=0.97, α=0.97, Δ=19, dropout supression=2



mixture is better than the single model:	0.47421	0.59736	0.53978	0.45982	0.58457	0
maximal absolute prediction error:	33.9325	13.7322	42.2052	42.1793	17.1499	22.9287
mean prediction error:	−0.038359	0.014908	−0.02256	−0.052713	0.012046	−0.0063408
median of prediction error:	0.010877	0.10452	0.023774	−0.044317	0.092584	−0.017578
std. dev. of prediction error:	3.2356	3.5536	3.5977	3.3209	3.5597	3.301
mean weights of the model:	0.43927	0.14132	0.049548	0.18743	0.17483	

U-dropout in h_1 between 200 and 1600, $(b_1, b_2) = (0.5, 1)$, $C = 10^1$

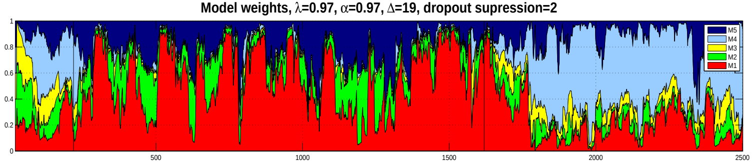
Model weights, λ=0.97, α=0.97, Δ=19, dropout supression=2



mixture is better than the single model:	0.47341	0.59856	0.54218	0.45702	0.58457	0
maximal absolute prediction error:	33.9325	13.7322	42.2052	42.1793	17.1499	22.9287
mean prediction error:	−0.038359	0.014908	−0.02256	−0.052713	0.012046	−0.0092456
median of prediction error:	0.010877	0.10452	0.023774	−0.044317	0.092584	−0.018714
std. dev. of prediction error:	3.2356	3.5536	3.5977	3.3209	3.5597	3.3047
mean weights of the model:	0.38646	0.13265	0.076847	0.2344	0.16204	

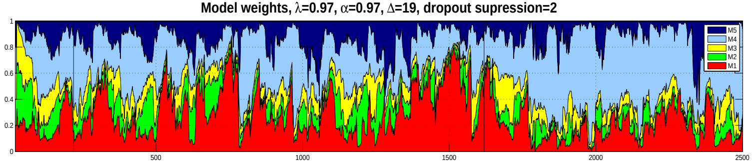
Experiments with the method IV

U-dropout in h_1 between 200 and 1600, $(b_1, b_2) = (0.9, 1)$, $C = 10^2$



mixture is better than the single model:	0.47181	0.59936	0.54098	0.45822	0.58697	0
maximal absolute prediction error:	33.9325	13.7322	42.2052	42.1793	17.1499	22.9287
mean prediction error:	−0.038359	0.014908	−0.02256	−0.052713	0.012046	−0.0058656
median of prediction error:	0.010877	0.10452	0.023774	−0.044317	0.092584	−0.023079
std. dev. of prediction error:	3.2356	3.5536	3.5977	3.3209	3.5597	3.3038
mean weights of the model:	0.41232	0.13696	0.064293	0.21052	0.16831	

U-dropout in h_1 between 200 and 1600, $(b_1, b_2) = (0.9, 1)$, $C = 10^1$



mixture is better than the single model:	0.47981	0.59696	0.54618	0.46261	0.57457	0
maximal absolute prediction error:	33.9325	13.7322	42.2052	42.1793	17.1499	22.9287
mean prediction error:	−0.038359	0.014908	−0.02256	−0.052713	0.012046	−0.030078
median of prediction error:	0.010877	0.10452	0.023774	−0.044317	0.092584	−0.0336
std. dev. of prediction error:	3.2356	3.5536	3.5977	3.3209	3.5597	3.3004
mean weights of the model:	0.26676	0.1129	0.10961	0.36787	0.13526	

Advantages:

- fast suppression of the model weights
- value of C can be used to adjust sensitivity to a_t^i
- during the model weights recovery, the affected models gather enough information for parameters and predictions

Disadvantages:

- slow recovery of the model weights
- direct manipulation with sufficient statistics
- heuristic

Modification of forgetting factor λ

Principle:

- significant decrease of the predictive variance (close to 0), setting mean value of regression coefficients to 0
- data update

$$\begin{array}{rcl} V_t^i & = & a_t^i \lambda V_{t-1}^i + \phi\phi' \\ \nu_t^i & = & a_t^i \lambda \nu_{t-1}^i + 1 \end{array} \rightarrow \begin{array}{rcl} \tilde{V}_t^i & = & V_t^i + (1 - a_t^i \lambda) V^{Ai} \\ \tilde{\nu}_t^i & = & \nu_t^i + (1 - a_t^i \lambda) \nu^{Ai} \end{array} \rightarrow f(\Theta | \tilde{V}_t^i, \tilde{\nu}_t^i)$$

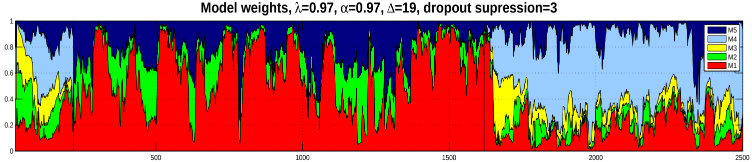
- alternative V^{Ai}, ν^{Ai} form a proper posterior pdf, keep the computations numerically stable and adjust sensitivity of the method

Effect:

- slower suppression of the respective $\omega_{t|t}^i$ if $a_t^i < 1$
- if $a_t^i = 1$, return to the regular operating state is very fast

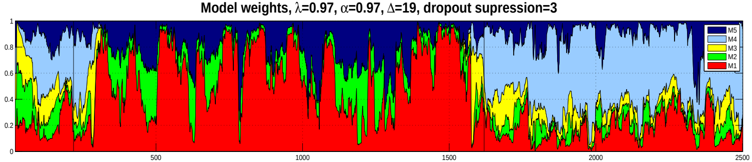
Experiments with the method I

U-dropout in h_1 between 200 and 1600, $(b_1, b_2) = (0, 1)$



mixture is better than the single model:	0.47221	0.59576	0.64854	0.60616	0.58297	0
maximal absolute prediction error:	33.9323	13.7319	42.2052	42.1793	17.0942	22.9287
mean prediction error:	−0.038359	0.014733	0.060504	0.018288	0.01197	−0.01024
median of prediction error:	0.010898	0.10451	0.10583	0.068812	0.087985	−0.011416
std. dev. of prediction error:	3.2356	3.5531	4.6013	4.5605	3.5586	3.3053
mean weights of the model:	0.42496	0.1391	0.055373	0.20023	0.17274	

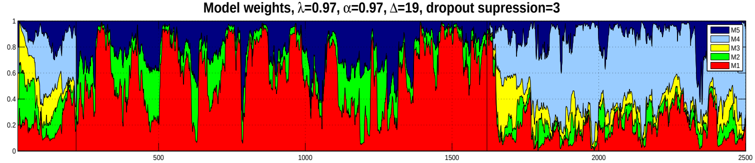
V-dropout in h_1 between 200 and 1600, $(b_1, b_2) = (0, 1)$



mixture is better than the single model:	0.46341	0.58737	0.64814	0.60496	0.57737	0
maximal absolute prediction error:	33.93226	13.7319	794.5426	531.1874	17.09418	22.92868
mean prediction error:	−0.038359	0.014733	0.19812	0.13551	0.01197	0.0027304
median of prediction error:	0.010898	0.10451	−0.068441	−0.12924	0.087985	−0.0090377
std. dev. of prediction error:	3.23564	3.55313	19.8175	14.3834	3.55865	3.34363
mean weights of the model:	0.39794	0.13586	0.067691	0.22433	0.16658	

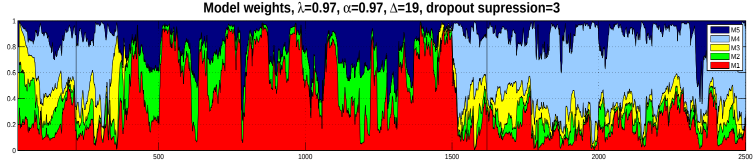
Experiments with the method II

U-dropout in h_1 between 200 and 1600, $(b_1, b_2) = (0.5, 1)$



mixture is better than the single model:	0.47261	0.59656	0.65694	0.61255	0.58377	0
maximal absolute prediction error:	33.9323	13.7319	42.2052	42.1793	17.0942	22.9287
mean prediction error:	−0.038359	0.014733	−0.1664	−0.1393	0.01197	−0.0080942
median of prediction error:	0.010898	0.10451	−0.102	−0.048443	0.087985	−0.013117
std. dev. of prediction error:	3.2356	3.5531	5.2597	5.1236	3.5586	3.3045
mean weights of the model:	0.42528	0.13919	0.055095	0.2001	0.17274	

V-dropout in h_1 between 200 and 1600, $(b_1, b_2) = (0.5, 1)$

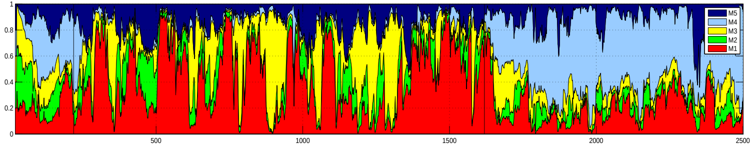


mixture is better than the single model:	0.45382	0.57857	0.62415	0.58337	0.56697	0
maximal absolute prediction error:	33.9323	13.7319	42.2052	42.1793	17.0942	22.9287
mean prediction error:	−0.038359	0.014733	−0.21036	−0.20424	0.01197	−0.0048862
median of prediction error:	0.010898	0.10451	−0.12391	−0.15508	0.087985	0.014699
std. dev. of prediction error:	3.2356	3.5531	4.6114	4.4006	3.5586	3.3708
mean weights of the model:	0.36207	0.13252	0.076646	0.25707	0.16409	

Experiments with the method III

U-dropout in h_1 between 200 and 1600, $(b_1, b_2) = (0.9, 1)$

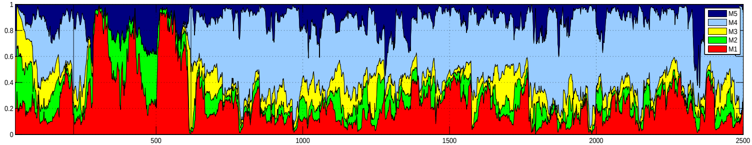
Model weights, λ=0.97, α=0.97, Δ=19, dropout supression=3



mixture is better than the single model:	0.43423	0.53099	0.62895	0.56497	0.52419	0
maximal absolute prediction error:	33.9323	13.7319	42.2052	42.1793	17.0942	22.9287
mean prediction error:	−0.038359	0.014733	−0.18761	−0.23454	0.01197	−0.087112
median of prediction error:	0.010898	0.10451	−0.12356	−0.20644	0.087985	−0.024079
std. dev. of prediction error:	3.2356	3.5531	4.1647	3.9704	3.5586	3.5287
mean weights of the model:	0.3257	0.11557	0.19341	0.22656	0.13116	

V-dropout in h_1 between 200 and 600, $(b_1, b_2) = (0.5, 1)$

Model weights, λ=0.97, α=0.97, Δ=19, dropout supression=3



mixture is better than the single model:	0.47341	0.58776	0.57777	0.4906	0.57097	0
maximal absolute prediction error:	33.9323	13.7319	42.2052	42.1793	17.0942	22.9287
mean prediction error:	−0.038359	0.014733	−0.054871	−0.0595	0.01197	−0.011883
median of prediction error:	0.010898	0.10451	0.081941	0.034518	0.087985	−0.013117
std. dev. of prediction error:	3.2356	3.5531	3.959	3.7219	3.5586	3.3132
mean weights of the model:	0.2464	0.10765	0.10329	0.40961	0.12546	

Advantages:

- faster reaction on a_t^i in the beginning and the end of the dropout
- statistics V^{Ai}, ν^{Ai} can be used to adjust sensitivity and reaction on status changes; even in the sense of the Method 2 (increase of predictive variance)

Disadvantages:

- different reaction on higher values of a_t^i than Method 2
- heuristic as well

Assigning weights to the data vector ϕ_t

Principle:

- significant decrease of the predictive variance (close to 0), setting mean value of regression coefficients to 0
- data are weighted by a_t^i
- data update

$$\begin{aligned} V_t^i &= \lambda V_{t-1}^i + (a_t^i)^2 \phi\phi' \\ \nu_t^i &= \lambda \nu_{t-1}^i + (a_t^i)^2 \end{aligned} \quad \rightarrow \quad \begin{aligned} \tilde{V}_t^i &= V_t^i + (1-\lambda) V^{Ai} \\ \tilde{\nu}_t^i &= \nu_t^i + (1-\lambda)\nu^{Ai} \end{aligned} \quad \rightarrow f(\Theta|\tilde{V}_t^i, \tilde{\nu}_t^i)$$

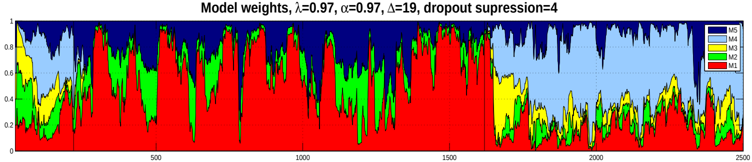
- alternative V^{Ai}, ν^{Ai} form a proper posterior pdf, keep the computations numerically stable and adjust sensitivity of the method

Effect:

- slower suppression of the respective $\omega_{t|t}^i$ if $a_t^i < 1$
- more tolerant
- if $a_t^i = 1$, return to the regular operating state is very fast

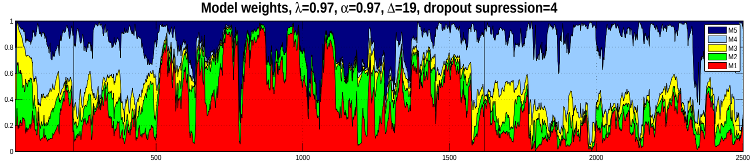
Experiments with the method I

U-dropout in h_1 between 200 and 1600, $(b_1, b_2) = (0, 1)$



mixture is better than the single model:	0.47101	0.59536	0.5006	0.44382	0.58457	0
maximal absolute prediction error:	33.9323	13.7319	42.2052	42.1793	17.0942	22.9287
mean prediction error:	−0.038359	0.014733	0.21971	0.20663	0.01197	−0.0080923
median of prediction error:	0.010898	0.10451	0.26095	0.21509	0.087985	−0.00046642
std. dev. of prediction error:	3.2356	3.5531	3.5137	3.4244	3.5586	3.3036
mean weights of the model:	0.42348	0.13849	0.057111	0.20179	0.17153	

V-dropout in h_1 between 200 and 1600, $(b_1, b_2) = (0, 1)$

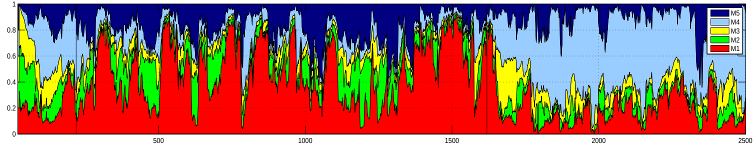


mixture is better than the single model:	0.47141	0.59776	0.54298	0.46301	0.58217	0
maximal absolute prediction error:	33.9323	13.7319	42.2052	42.1793	17.0942	22.9287
mean prediction error:	−0.038359	0.014733	−0.046334	−0.068999	0.01197	−0.010759
median of prediction error:	0.010898	0.10451	−0.0083144	−0.048442	0.087985	−0.011168
std. dev. of prediction error:	3.2356	3.5531	3.5915	3.3206	3.5586	3.3038
mean weights of the model:	0.33472	0.12484	0.08561	0.29215	0.15508	

Experiments with the method II

U-dropout in h_1 between 200 and 1600, $(b_1, b_2) = (0.5, 1)$

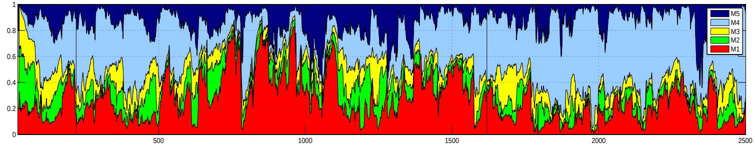
Model weights, λ=0.97, α=0.97, Δ=19, dropout supression=4



mixture is better than the single model:	0.47101	0.60016	0.54218	0.45342	0.58257	0
maximal absolute prediction error:	33.9323	13.7319	42.2052	42.1793	17.0942	22.9287
mean prediction error:	−0.038359	0.014733	−0.019114	−0.047203	0.01197	−0.011833
median of prediction error:	0.010898	0.10451	0.024725	−0.044248	0.087985	−0.0077297
std. dev. of prediction error:	3.2356	3.5531	3.5908	3.314	3.5586	3.3021
mean weights of the model:	0.35583	0.12715	0.094031	0.26055	0.15484	

V-dropout in h_1 between 200 and 1600, $(b_1, b_2) = (0.5, 1)$

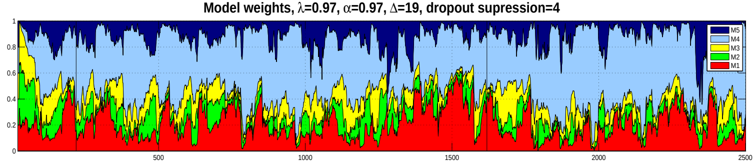
Model weights, λ=0.97, α=0.97, Δ=19, dropout supression=4



mixture is better than the single model:	0.47901	0.59736	0.54138	0.46341	0.57337	0
maximal absolute prediction error:	33.9323	13.7319	42.2052	42.1793	17.0942	22.9287
mean prediction error:	−0.038359	0.014733	−0.026649	−0.055947	0.01197	−0.023684
median of prediction error:	0.010898	0.10451	0.01524	−0.03961	0.087985	−0.020248
std. dev. of prediction error:	3.2356	3.5531	3.5958	3.3197	3.5586	3.2991
mean weights of the model:	0.26423	0.11138	0.10983	0.37122	0.13574	

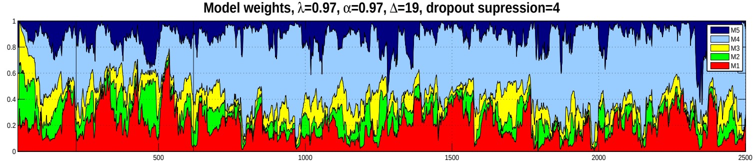
Experiments with the method III

U-dropout in h_1 between 200 and 1600, $(b_1, b_2) = (0.9, 1)$



mixture is better than the single model:	0.48061	0.59616	0.54658	0.46461	0.57297	0
maximal absolute prediction error:	33.9323	13.7319	42.2052	42.1793	17.0942	22.9287
mean prediction error:	−0.038359	0.014733	−0.022169	−0.051919	0.01197	−0.037832
median of prediction error:	0.010898	0.10451	0.024725	−0.044279	0.087985	−0.034785
std. dev. of prediction error:	3.2356	3.5531	3.5961	3.3195	3.5586	3.2933
mean weights of the model:	0.20696	0.10096	0.11592	0.44884	0.11972	

V-dropout in h_1 between 200 and 600, $(b_1, b_2) = (0.5, 1)$



mixture is better than the single model:	0.47741	0.59696	0.54458	0.46261	0.57257	0
maximal absolute prediction error:	33.9323	13.7319	42.2052	42.1793	17.0942	22.9287
mean prediction error:	−0.038359	0.014733	−0.022723	−0.053528	0.01197	−0.033337
median of prediction error:	0.010898	0.10451	0.024725	−0.037331	0.087985	−0.022616
std. dev. of prediction error:	3.2356	3.5531	3.5969	3.3211	3.5586	3.295
mean weights of the model:	0.20602	0.10242	0.11408	0.44992	0.11996	

Advantages:

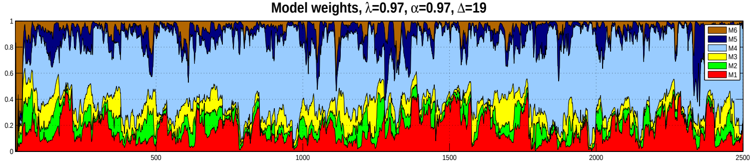
- theoretically based
- fast reactions

Disadvantages:

- maybe less sensitive?

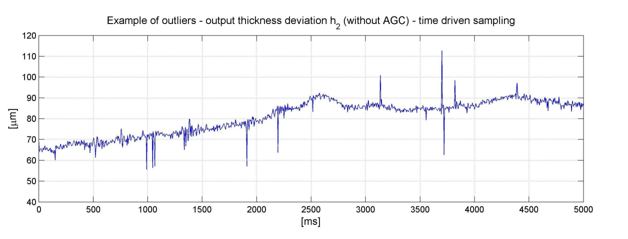
Another interesting issue

No dropout, 6^{th} model containts only one data channel with a sequence -1, 1, -1, 1, -1, . . . , i.e. pure nonsense

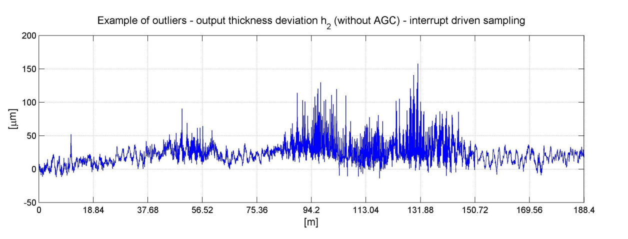


mixture is better than the single model:	0.4954	0.59936	0.56018	0.47661	0.57417	0.47661	0
maximal absolute prediction error:	33.9325	13.7322	42.2052	42.1793	17.1499	11.2315	19.4021
mean prediction error:	−0.038359	0.014908	−0.02256	−0.052713	0.012046	−0.0077814	−0.057061
median of prediction error:	0.010877	0.10452	0.023774	−0.044317	0.092584	0.11082	−0.038367
std. dev. of prediction error:	3.2356	3.5536	3.5977	3.3209	3.5597	3.0919	3.2403
mean weights of the model:	0.16623	0.080495	0.099302	0.48798	0.10139	0.057005	

Particular corruption of data



isolated outliers



dirty strip

- corrected wrong likelihood computation $\rightarrow$ method 1 takes at least small effect
- "emergency" artificial decreasing of D_{11} (methods 3 and 4) yields faster recovery than increasing of D_{11} (method 2)
- methods 3 and 4 can potentially increase D_{11}, too, if alternative statistic V^A has big entries
- nonsense model still admitted
- adaptive median filter in progress
- data scaling and filtration of outliers seems to be necessary

Thank you for your attention

questions and comments are welcome