# Receding Horizon Prediction by Bayesian Combination of Multiple Predictors

Ståhl, Fredrik; Johansson, Rolf

*Published in:*
Proc. 51st IEEE Conf. Decision and Control (CDC 2012), December 10-13, 2012. Maui, Hawaii, USA

2012

51st IEEE Conference on Decision and Control
December 10-13, 2012. Maui, Hawaii, USA

# Receding Horizon Prediction by Bayesian Combination of Multiple Predictors

F. Ståhl and R. Johansson

***Abstract*—This paper presents a novel online approach of merging multiple different predictors of time-varying dynamics into a single optimized prediction. Different predictors are merged by recursive weighting into a single prediction using regularized optimization. The approach is evaluated on two different cases of data with shifting dynamics; one example of prediction using several approximate models of a linear system and one case of glucose prediction on a non-linear physiologically based simulated type I diabetes data using several parallel linear predictors. The performance of the combined prediction significantly reduced the total prediction error compared to each predictor in each example.**

## I. Introduction

In many applications in science and technology the prediction of future values for some data series is of interest. For this purpose different models can be identified and utilized to produce predictions. When more than one model is constructed the question arises whether it is more useful to use one of them solely, or if it is possible to gain additional prediction accuracy by combining their outcomes. Accuracy may be gained from merging due to mismodeling or to changing dynamics in the underlying data creating process. Of special interest are time-varying systems with unknown, or complex, shifts in dynamics, where a single model capturing the system behaviour may be infeasible, e.g., for practical identification concerns.

Merging models for the purpose of prediction has been developed in different research communities. In the meteorological and econometric communities regression-oriented ensemble prediction has been a vivid research area since the late ’60s, see e.g. [18] and [8].

Also in the machine learning community the question of how different predictors or classifiers can be used together for increased performance has been investigated and different algorithms developed such as the bagging, boosting [5] and weighted majority [15] algorithms, and online versions of these [13], [17].

In most approaches the merged prediction $\hat{y}^e_k$ at time $k$ is formed by a linear weighted average of the individual predictors $\hat{\mathbf{y}}_k$.

$$\hat{y}^e_k = \mathbf{w}_k^{\mathbf{T}} \hat{\mathbf{y}}_k \qquad (1)$$

It is also common to restrict the weights $\mathbf{w}_k$ to $[0,1]$. The possible reasons for this are several, where the interpretation of the weights as probabilities, or rather Bayesian beliefs, is the dominating. Such restrictions are however not always applicable, e.g. in the related optimal portfolio selection problem where negative weight (short selling) can reduce the portfolio risk [9].

A special case considering distinct switches between different linear system dynamics has been studied mainly in the control community. The data stream and the underlying dynamic system are modelled by pure switching between different filters derived from these models, i.e., the weights $w_k$ can only take value 1 or 0. A lot of attention has been given to reconstructing the switching sequence, see e.g. [10], [16]. From a prediction viewpoint the current dynamic mode is of primary interest, and it may suffice to reconstruct the dynamic mode for a limited section of the most recent time points in a receding horizon fashion [2].

Combinations of specifically adaptive filters has also stirred some interest in the signal processing community. Typically, filters with different update pace are merged to benefit from each filter’s specific change responsiveness respectively steady state behaviour [3].

Finally, in fuzzy modeling, soft switching between multiple models is offered using fuzzy membership rules in the Takagi-Sugeno systems [20].

This paper presents a novel approach combining elements from both switching and averaging techniques above, forming a ’soft’ switcher in a Bayesian framework. The paper is organized as follows; in Sec. II the problem formulation is presented, Sec. III describes the algorithm, in Sec. IV and Sec. V results from two example are given and Sec. VI concludes the paper.

## II. Problem Formulation

Considering the above, this paper addresses the following problem: A non-stationary data stream $z_k : \{y_k, u_k\}$ arrives with a fixed sample rate, set to 1 for notational convenience, at time $t_k \in \{1,2,...\}$. The data stream contains a variable of primary interest called $y_k \in \mathbb{R}$ and additional variables $u_k$. The data stream can be divided into different periods $T_{S_i}$ of similar dynamics $S_i \in S = [1...n]$, and where $s_k \in S$ indicates the current dynamic mode at time $t_k$. The system changes between these different modes according to unknown dynamics.

Given $m$ number of expert $p$-steps-ahead predictions at time $t_k$, $\hat{y}^j_{k+p|k}, j \in \{1,..m\}$, each utilizing different methods, and/or different training sets; how is an optimal $p$-steps-ahead prediction $\hat{y}^e_{k+p|k}$ of the primary variable $y(t_k)$, using a predefined norm and under time-varying conditions, determined?

F. Ståhl and R. Johansson are with Dept. Automatic Control, Lund University, PO Box 118, SE22100 Lund Sweden, Fredrik.Stahl@control.lth.se, Phone +4646 222 87 84, Fax +4646 138118

978-1-4673-2064-1/12/$31.00 ©2012 IEEE

## III. Sliding Window Bayesian Model Averaging

Apart from conceptual differences between the different approaches to ensemble prediction the most important difference is how the weights are determined. Numerous different methods exist, ranging from heuristic algorithms [3], [20] to theory based approaches, e.g. [11]. Specifically, in a Bayesian Model Averaging framework [11], which will be adopted in this paper, the weights are interpreted as partial beliefs in each predictor $M_i$, and the merging is formulated as:

$$p(y_{k+p}|D_k) = \sum_i p(y_{k+p}|M_i, D_k)p(M_i|D_k) \tag{2}$$

and if only point-estimates are available one can e.g. use:

$$\begin{aligned} \hat{y}^e_{k+p|k} &= \mathbb{E}(y_{k+p}|D_k) && (3)\\ &= \sum_i \mathbb{E}(M_i|D_k)\mathbb{E}(y_{k+p}|M_i, D_k) && (4)\\ &= \mathbf{w}_k^T \hat{\mathbf{y}}_k && (5)\\ \mathbf{w}_k^{(i)} &= \mathbb{E}(M_i|D_k) && (6)\\ p(\mathbf{w}_k^{(i)}) &= p(M_i|D_k) && (7) \end{aligned}$$

Where $\hat{y}^e_{k+p}$ is the combined prediction of $y_{k+p}$ using information available at time $k$, $D_k : \{z_{1:k}\}$ is the data received up until time $k$, $\mathbf{w}_k^{(i)}$ indicates position $i$ in the weight vector. The conditional probability of predictor $M_i$ can be further expanded by introducing the latent variable $\Theta_j$.

$$p(M_i|D_k) = \sum_j p(M_i|\Theta_j, D_k)p(\Theta_j|D_k) \tag{8}$$

or in matrix notation

$$\mathbf{p}(\mathbf{w}_k) = \left[\mathbf{p}(\mathbf{w}_{k|\theta_k=1}) \ldots \mathbf{p}(\mathbf{w}_{k|\theta_k=n})\right] \mathbf{p}(\Theta|D_k) \tag{9}$$

Here $\Theta_j$ represents a *predictor mode* in a similar sense to the dynamic mode $S_j$, and likewise $\theta_k$ the prediction mode at time $k$. $\mathbf{p}(\Theta|D_k)$ is a column vector of $p(\Theta_j|D_k), j = \{1 \ldots m\}$ and $\mathbf{p}(\mathbf{w}_{k|\Theta_i})$ is a row vector of the joint prior distribution of the conditional weights of each predictor model given the predictor mode $\Theta_i$.

Data for estimating the distribution for $\mathbf{p}(\mathbf{w}_{k|\Theta_i})$ is based on labelled training data used in the following constrained optimization.

$$\{\mathbf{w}_{k|\Theta_i}\}_{T_{\Theta_i}} = \arg\min \sum_{i=k-N/2}^{k+N/2} \mathscr{L}(y(t_i), \mathbf{w}^T_{k|\Theta_i}\hat{\mathbf{y}}_{\mathbf{i}}), \quad k \in T_{\Theta_i} \tag{10}$$

s.t. $\sum_j w^{(j)}_{k|\Theta_i} = 1$.
where $T_{S_i}$ represents the time points corresponding to dynamic mode $S_i$, $N$ is the size of the evaluation window, $\mathscr{L}(y, \hat{y})$ is a cost function. From these data sets the prior distributions can be estimated by the Parzen window method [4], giving mean $\mathbf{w}_{0|P_i} = \mathbb{E}(\mathbf{w}_{k|\Theta_i})$ and covariance matrix $\mathbf{R}_{\Theta_i}$. An alternative to the Parzen approximation is of course to estimate a more parsimoniously parametrized pdf (e.g., Gaussian) for the extracted data points.

Now, in each time step $k$ the $\mathbf{w}_{k|\theta_{k-1}}$ is determined from the sliding window optimization below, using the current active mode $\theta_{k-1}$. For reasons soon explained, only $\mathbf{w}_{k|\theta_{k-1}}$ is thus calculated.

$$\begin{aligned} \mathbf{w}_{k|\theta_{k-1}} = \arg\min & \sum_{j=k-N}^{k-1} \mu^{k-j} \mathscr{L}(y_j, \mathbf{w}^T_{k|\theta_{k-1}}\hat{\mathbf{y}}_j) \qquad (11)\\ & + (\mathbf{w}_{k|\theta_{k-1}} - \mathbf{w}_{0|\theta_{k-1}})\Lambda(\mathbf{w}_{k|\theta_{k-1}} - \mathbf{w}_{0|\theta_{k-1}})^T\\ \text{s.t.} & \sum_j w^{(j)}_{k|\theta_{k-1}} = 1 \end{aligned}$$

Here, $\mu_j$ is a forgetting factor, and $\Lambda$ is the regularization matrix. Now, to infer the posterior $\mathbf{p}(\Theta|D_k)$ it would normally be natural to set this probability function equal to the corresponding posterior pdf for the dynamic mode $\mathbf{p}(S|D_k)$. However, problems arise if $\mathbf{p}(S|D_k)$ is not directly possible to estimate from the dataset $D_k$. This is circumvented by using the information provided by the $\mathbf{p}(\mathbf{w}_{k|\theta_k})$ estimated from the data retrieved from equation (10) above. The $\mathbf{p}(\mathbf{w}_{k|\theta_k})$ prior density functions can be seen as defining the region of validity for each predictor mode. If the $\mathbf{w}_{k|\theta_{k-1}}$ estimate leaves the current active mode region $\theta_{k-1}$ (in a sense that $\mathbf{p}(\mathbf{w}_{k|\theta_{k-1}})$ is very low) it can thus be seen as an indication of that a mode switch has taken place. A logical test is used to determine if a mode switch has occurred. The predictor mode is switched to mode $\Theta_i$, if:

$$\begin{cases} p(\Theta_i|\mathbf{w}_k, D_k) > \lambda, \text{ and} \\ p(\mathbf{w}_k|\Theta_i, D_k) > \delta \end{cases} \tag{12}$$

where

$$p(\Theta_i|\mathbf{w}_k, D_k) = \frac{p(\mathbf{w}_k|\Theta_i, D_k)p(\Theta_i|D_k)}{\sum_j p(\mathbf{w}_k|\Theta_j, D_k)p(\Theta_j|D_k)} \tag{13}$$

Where a $\lambda$ somewhat larger than 0.5 gives a hysteresis effect to avoid chattering between modes, and $\delta$ assures that non-conclusive situations, evaluated on the outskirts of the probability functions, don't result in switching. Unless otherwise estimated from data, the conditional probability of each prediction mode $p(\Theta_i|D_k)$ is set equal for all possible modes, and thus cancels in (13). The logical test is evaluated using the priors received from the pdf estimate and the $\mathbf{w}_{\mathbf{k}|\theta_\mathbf{k}}$ received from (11). If a mode switch is considered to have occurred, (11) is rerun using the new predictor mode.

Now, since only one prediction mode $\theta_k$ is active; (9) reduces to $\mathbf{p}(w_k) = \mathbf{p}(w_{k|\theta_k})$.

### A. Choice of $\mathscr{L}$

Cost function should be chosen by the specific application in mind. A natural choice for interpolation is the 2-norm, but in certain situations asymmetric cost functions are more appropriate.

### B. Parameter choice

The length $N$ of the evaluation period is, together with the forgetting factor $\mu$, a crucial factor determining how fast the ensemble prediction reacts to sudden changes in dynamics. A small forgetting factor will put much emphasis on recent data making it more agile to sudden changes. However, the drawback is of course that noise sensitivity increases.

$\Lambda$ should also be chosen such that a sound balance between flexibility and robustness is found, i.e., a too small $\Lambda$ may result in over-switching, whereas a too large $\Lambda$ will give a stiff and inflexible predictor. Furthermore, $\Lambda$ should force the weights to move within the perimeter defined by $p(\mathbf{w}|\Theta_i)$. This is approximately accomplished by setting $\Lambda$ equal to the inverse of the covariance matrix $\mathbf{R}_{\theta_k}$, thus representing the pdf as a Gaussian distribution in the regularization.

### C. Nominal mode

Apart from the estimated prediction mode centres an additional predictor mode can be added corresponding to a heuristic fall-back mode. In the case of sensor failure or other situations where loss of confidence in the estimated predictor modes arises, each predictor may seem equally valid. In this case a fall-back mode to resort to may be the equal weighting. This is also a natural start for the algorithm. For these reasons a nominal mode $p(\bar{\mathbf{w}}^{\mathbf{0}}) \in N(\mathbf{1/n}, \mathbf{I})$ is added to the set of predictor modes.

Summary of algorithm

1) Estimate $n$ number of predictors according to best practice.
2) Run the constrained estimation (10) on labelled training data and retrieve the sequence of $\{\mathbf{w}_{k|\Theta_i}\}_{T_{\theta_i}}, \forall i \in \{1,..,n\}$.
3) Classify different predictor modes and determine density functions $\mathbf{p}(\mathbf{w}_{k|\Theta_i})$ for each mode $S_i$ from the training results by supervised learning. If possible; estimate $p(S|D)$.
4) Initialize mode to the nominal mode.
5) For each time step; calculate $\mathbf{w_k}$ according to (11).
6) Test if switching should take place by evaluating (12) and (13), and switch predictor mode if necessary and recalculate new $\mathbf{w}_k$ according to (11).
7) Go to 5).

## IV. Example: Prediction using approximate lower-order models

### A. Data

Data were generated using a switched fourth-order ARX system, where the A-polynomial switches between three different models $M_A, M_B, M_C$, with poles according to Table I. The B-polynomial was simply a one step delay, and white noise $N(0, 0.25)$ was added to the output channel. A PRBS signal was used for input.

The active dynamic mode $s_k \in S$ switches according to a transition matrix $M$ between dynamic mode A,B and C.

$$s_{k+1} = Hs_k \tag{14}$$

$$H = \begin{bmatrix} 0.99 & 0.005 & 0.005 \\ 0.005 & 0.99 & 0.005 \\ 0.005 & 0.005 & 0.99 \end{bmatrix} \tag{15}$$

TABLE I
Poles of the data generating processes.

| Model | Poles |
|---|---|
| $M_A$ | $0.8, 0.1, -0.3 + i\sqrt{0.41}, -0.3 - i\sqrt{0.41}$ |
| $M_B$ | $0.9, 0.2, -0.2, -0.5$ |
| $M_C$ | $0.8, -0.2, -0.4, -0.4$ |

A labelled training set of 2000 samples and a 2000 sample validation set were simulated in 40 different batches. An example of a training data set can be seen in Fig. 1.

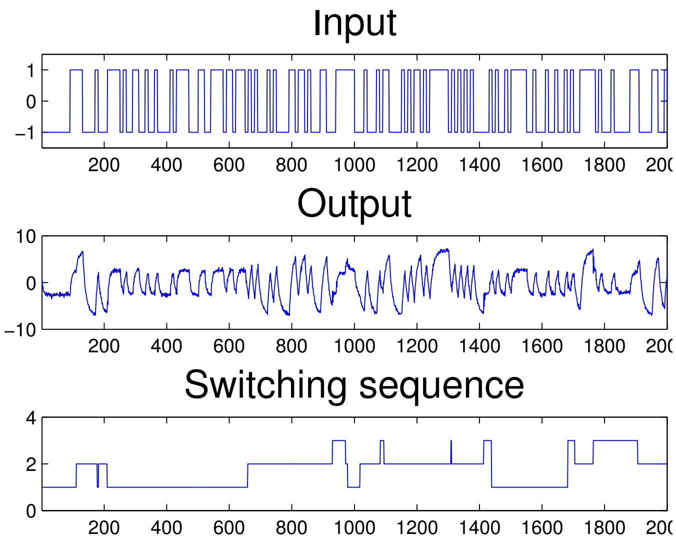

Fig. 1. Training Data. Upper plot: input, middle plot: output and lower plot: switching sequence of dynamic mode.

### B. Predictors

To simulate modeling errors three prediction models $M_I - M_{III}$ were set up as reduced order approximations of the corresponding state-space models of the data generating processes. Model reduction was undertaken by singular value evaluation to the second order [12]. Using these models and their associated Kalman filters 50 step prediction length was evaluated.

### C. Cost function

For this example the 2-norm was used.

### D. Parameter Choices

Different values for the tunable parameters $N$ and $\mu$ were evaluated. 20 batches for the combinations of $\{10, 20, 30\}$ and $\{0.8, 0.9, 1\}$, and 20 batches for the combination of $\{25, 50, 75\}$ and $\{0.7, 0.8, 0.9\}$. The parameters $\lambda$ and $\delta$ for the switching test were fixed to 0.6 and $3 \cdot 10^{-3}$.

### E. Evaluation Metric

To evaluate the predictive performance, the squared sum of prediction errors was compared to the squared sum of prediction errors using a pure switching strategy where it (optimally) has been assumed that the dynamic mode at the time of prediction was known.

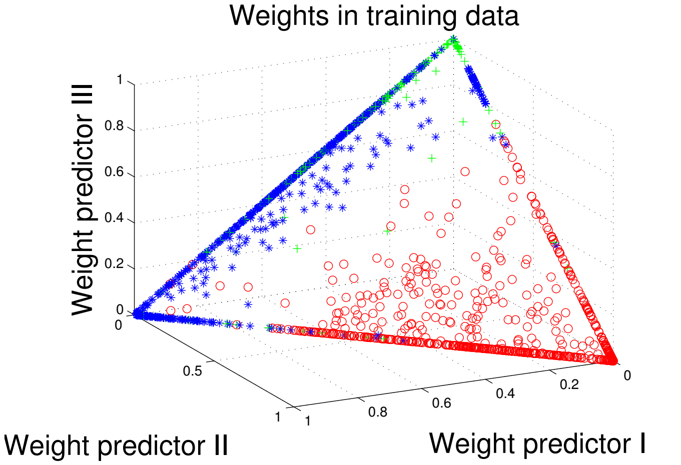

Fig. 2. Distribution of weights in the training data retrieved by (10). Blue stars represents $t_k \in T_A$ for Mode A, Red circles: $T_B$ for Mode B and Green crosses: $T_C$ for Mode C.

The corresponding probability distribution for each mode, projected onto the $(w_1, w_2)$-plane, estimated by Parzen window technique, can be seen in Fig. 3 together with the pdf of the nominal mode. The densities have higher values in the corners $[1,0,0]$, $[0,1,0]$ and $[0,0,1]$, but with means $\mathbf{w_{0|1}} = [0.57, 0.03, 0.4]$, $\mathbf{w_{0|2}} = [0.18, 0.76, 0.06]$ and $\mathbf{w_{0|3}} = [0.25, 0.03, 0.72]$ defining the expected weights for each predictor mode.

### *F. Results*

*1) Training the mode switcher:* Using the labelled training set the pdf:s $\mathbf{p}(\mathbf{w}_{k|\Theta_i})$ were estimated for each batch using the different $N$ values. For this example the best evaluation record length for the estimation task was 10. In Fig. 2 an example of the distribution of $\{\mathbf{w}_{k|\Theta_i}\}_{S_i}$ along the $\{w_1 + w_2 + w_3 = 1, 0 \leq w_i \leq 1\}$ plane can be seen for one representative training batch.

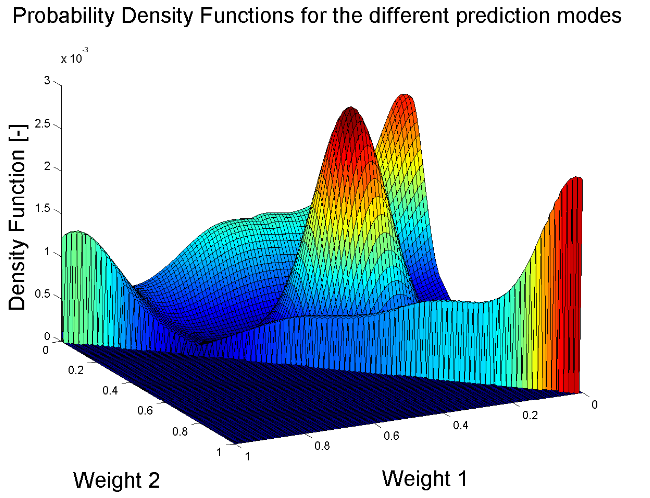

Fig. 3. Estimated probability density functions for the weights in the training data, including the nominal mode.

*2) Evaluation of parameter choices:* Comparing the predictive performance for the different value combinations of $N$ and $\mu$, the slightly better choice over the others was $[25, 0.8]$. Table II summarizes the predictive performance for each combination of $N$ and $\mu$.

TABLE II
SUMMARY OF PREDICTIVE PERFORMANCE USING DIFFERENT $N$ AND $\mu$ ON VALIDATION DATA OVER ALL SIMULATED BATCHES, EVALUATED AS MEAN $\frac{\sum e^2_{N_i,\mu_j}}{\sum e^2_{opt}}$, WHERE $e_{N_i,\mu_j}$ IS CORRESPONDING PREDICTION ERROR(PE) AND $e_{opt}$ IS THE PE WHEN USING THE OPTIMAL SWITCHING STRATEGY.

| $N_i/\mu_j$ | 1 | 0.9 | 0.8 | 0.7 |
|---|---|---|---|---|
| 10 | 0.92 | 0.91 | 0.90 | - |
| 20 | 0.95 | 0.94 | 0.92 | - |
| 25 | - | 0.91 | **0.87** | 0.87 |
| 30 | 0.94 | 0.92 | 0.88 | - |
| 50 | - | 1.05 | 1.04 | 1.01 |
| 75 | - | 0.95 | 0.90 | 0.88 |

*3) Predictive performance:* The merged prediction was compared on the validation data, using the best choices of $N = 25$ and $\mu = 0.8$, to 1) each individual predictor, 2) an unregularized version of (11) without switching functionality, and 3) to the optimal pure switching strategy. The results are summarized in Table III.

TABLE III
SUMMARY OF PREDICTIVE PERFORMANCE ON VALIDATION DATA OVER ALL SIMULATED BATCHES.

| Predictor | $\frac{\sum e^2}{\sum e^2_{opt}}$ |
|---|---|
| Predictor I | 1.07 |
| Predictor II | 2.76 |
| Predictor III | 1.39 |
| Merged Predictor | 0.87 |
| Unregularized Merged Predictor | 0.93 |
| Optimally Switched Predictor | 1.0 |

Compared to the other approaches a 7% improvement can be seen to the unregularized version, and a 13% improvement to the optimal switching scheme.

Looking at the distribution of the weights for the validation data in Fig. 4 it's apparent that the merging mechanism has concentrated these around the prediction mode centres, especially if comparing to the corresponding distribution for the unregularized version, see Fig. 5.

Switching between the different prediction modes in comparison to the dynamic mode for the validation data can be seen in Fig 6 for a representative batch.

### *G. Discussion*

*1) Parameter Choice:* The optimal choices of $N$ and $\mu$, are unsurprisingly, closely connected. These parameters must be set with the specific dynamics in mind, and are probably difficult to determine beforehand. $\lambda$ and $\delta$ should probably not be set too low in order to avoid uncalled for switching,

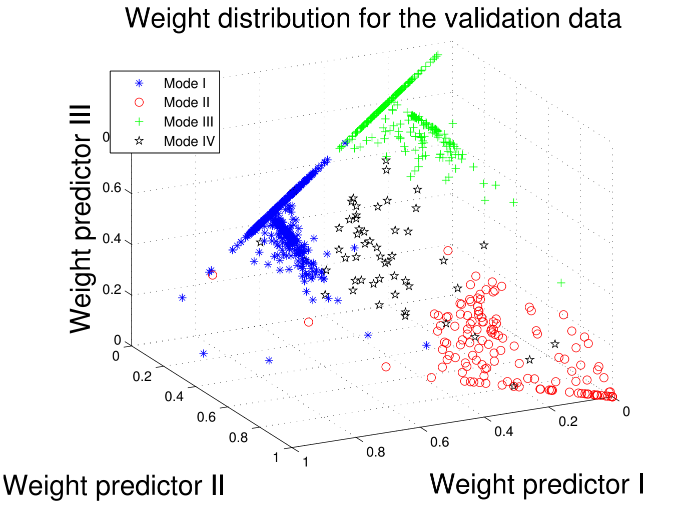

Fig. 4. Distribution of weights in the test data using the estimated pdf:s and expected weights.

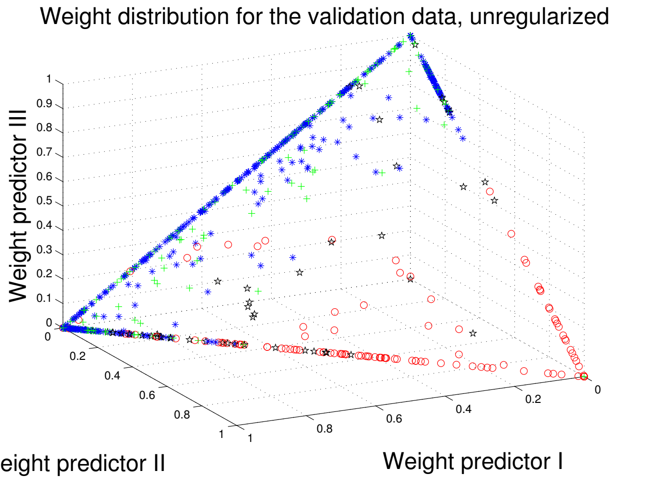

Fig. 5. Distribution of weights in the test data using the unregularized merging predictor.

and the values used are deemed correct from this aspect. The regularization penalty term $\mathbf{R}$ is on the other hand an important parameter determining the sensitivity to switching and robustness to noise. Analysis of tuning aspects of this parameter has however been left out for space limit reasons.

*2) Predictive Performance:* The merged predictor clearly outperformed each of the individual predictors, and also the unregularized version as well as the optimal pure switching predictor. The latter can be explained by that the merged predictor offers some extra robustness to sudden dynamical changes, as all predictors to some extent are used in all situations. The unregularized version has quite good performance, but the regularization in the proposed merging mechanism reduces the impact of noise making it slightly better.

Regarding the distribution of weights and the corresponding prior pdf:s it is interesting to notice that the merging mechanism puts almost equal confidence in predictor I and III for dynamic mode $A$. This could be interpreted as a means to handle and improve the predictive quality given the modeling errors in model $M_I$ in comparison to $M_A$.

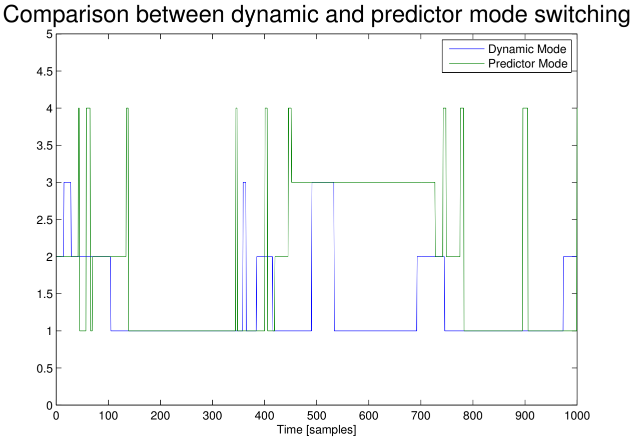

Fig. 6. Example of switching between different predictor modes in the validation data. Here predictor mode 4 represents the nominal mode.

## V. Example: Diabetes Type I Blood Glucose Prediction

### A. *Insulin Dependent Diabetes Mellitus*

Insulin Dependent Diabetes Mellitus (IDDM) is a chronic metabolic disease characterized by impaired plasma glucose regulation. Maintaining normoglycemia is crucial in order to avoid both immediate and long-term complications, and to this end, several insulin injection are taken daily. To facilitate the management of this therapy, blood glucose (BG) is measured using different devices, among them the so-called Continuous Glucose Measurement System (CGM) e.g., [1], which provides glucose values every 5 or 10 min. Accurate predictions of near-time glucose evolution would improve means to tight glucose control and further improve life-quality for these patients.

### B. *Data*

Data was generated using the non-linear metabolic simulation model described in [7] with parameter values obtained from the authors.

Twenty datasets, each corresponding to 8 days, were generated. Two dynamic modes $A$ and $B$ were simulated by, after 4 days, changing four model parameters (following the notation in [7]) $k_1, k_i$, $k_{p3}$ and $p_{2u}$, related to endogenous glucose production and insulin and glucose utilization. One data set was used for training and the others were considered test data.

A section of four days, including the period when the dynamic change takes place, of the training data and an example test data can be seen in Fig. 7.

Timing and size of meals were generated with some normal randomization for each data set according to Table IV. The amount of insulin administered for each meal was also randomized by normally distributed noise with a 20 % standard deviation.

TABLE IV
MEAL AMOUNT AND TIMING RANDOMIZATION. STANDARD DEVIATION IN PARENTHESIS.

| Meal | Time | Amount carbohydrates (g) |
|---|---|---|
| Breakfast | 08:00 (30 min) | 45 (5) |
| Lunch | 12:30 (30 min) | 70 (10) |
| Dinner | 19:00 (30 min) | 80 (10) |

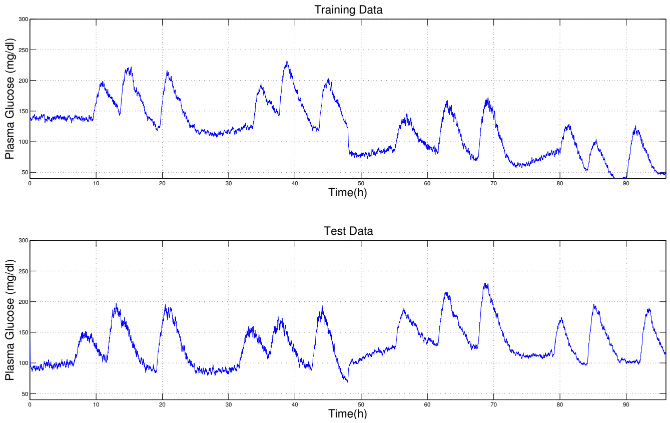

Fig. 7. Middle four days of the plasma glucose Training and Test Data.

Noise was added by perturbing some crucial model parameters $p_i$ in each simulation step; $p_i(t) = (1+\delta(t))p_i^0$, where $p_i^0$ represent nominal value and $\delta(t) \in N(0, 0.2)$. The affected parameters are (again following the notation in [7]) $k_1, k_2, p_{2u}, k_i, m_1, m_{30}, m_2, k_{sc}$.

### C. Predictors

Three models, based on subspace based technique, were identified using the N4SID algorithm of the Matlab System Identification Toolbox. Model order $[2-4]$ was determined by the Akaike criterion [12]. The first model $I$ was estimated using data from dynamic mode A in the training data, and the third $III$ from the mode $B$ data, and the final model $II$ from the entire training data set. Thus, model $I$ and $III$ are each specialized, whereas $II$ is an average of the two dynamic modes. The models were evaluated for a prediction horizon of 60 min.

### D. Cost function

A suitable cost function for determining appropriate weights should take into account that the consequences of acting on too high glucose predictions in the lower BG region (<90 mg/dl) could possibly be life threatening. The margins to low blood glucose levels that may result in coma and death are small, and blood glucose levels may fall rapidly. Hence, much emphasis should be put on securing small positive predictive errors and sufficient time margins for alarms to be raised in due time in this region. In the normoglycemic region (here defined as 90-200 mg/dl) the predictive quality is of less importance. This is the glucose range that non-diabetics normally experience, and thus can be considered, from a clinical viewpoint in regards to possible complications, a safe region. However, due to the possibility of rapid fluctuation of the glucose into unsafe regions some considerations of predictive quality should be maintained.

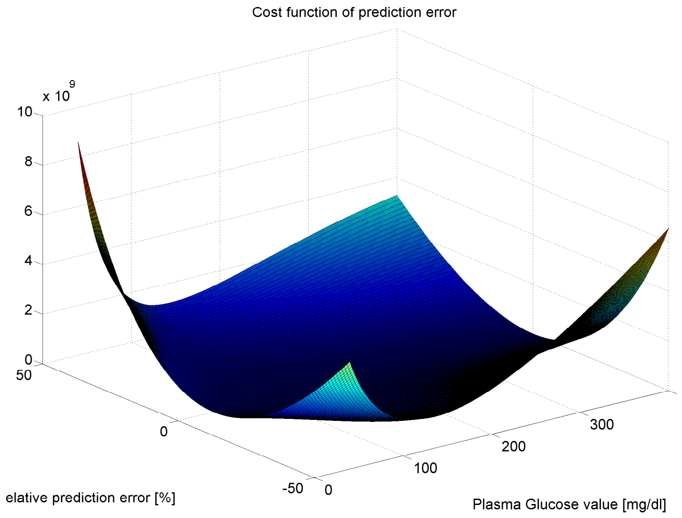

Fig. 8. Cost function of relative prediction error.

Based on the cost function in [14] the selected cost function incorporates these features; asymmetrically increasing cost of the prediction error depending on the absolute glucose value and the sign of the prediction error.

In Fig. 8 the cost function can be seen plotted against relative prediction error and absolute blood glucose value.

*1) Correspondence to the Clarke Grid Error Plot:* A de facto accepted standardized metric of measuring the performance of CGM signals in relation to reference measurements, and often used to evaluate glucose predictors, is the Clarke Grid Plot [6]. This metric meets the minimum criteria raised earlier. However, other aspects makes it less suitable; no distinction between prediction errors within error zones, instantaneous switches in evaluation score, etc.

In Fig. 9 the isometric cost of the chose cost function for different prediction errors at different BG values has been plotted together with the Clarke Grid Plot. The boundaries of the A/B/C/D/E areas of the Clarke Grid can be regarded lines of isometric cost according to the Clarke metric. In the figure the isometric cost of the cost function has been chosen to correspond to the lower edge defined by the intersection of the A and B Clarke areas. Thus, the area enveloped by the isometric cost can be regarded as the corresponding A area of this cost function. Apparently it puts much tougher demands both in the lower and upper BG regions in comparison to the Clarke Plot.

### E. Parameter Choice

$\mu$ was set to 0.8 and $N$ to 40 minutes.

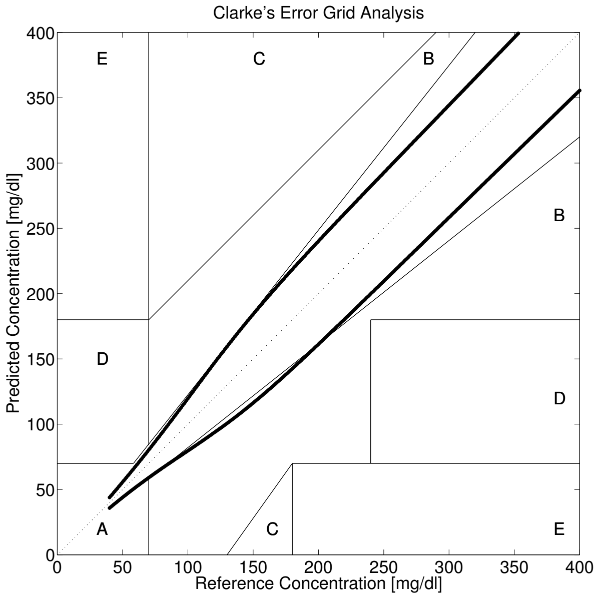

Fig. 9. Isometric cost in comparison to the Clarke Grid.

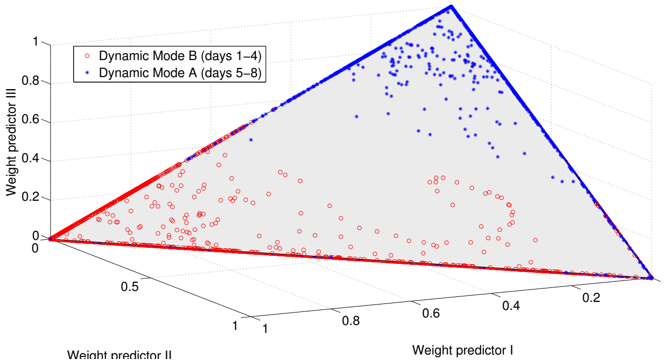

Fig. 10. Distribution of weights in the training data using the unconstrained ensemble version.

### F. Results

*1) Training the mode switcher:* The three predictors were used to create three sets of 60 minute ahead predictions for the training data. Using (10) with $N = 10$ the weights $\mathbf{w}_k$ were determined. In Fig. 10 the distribution of the $\mathbf{w}_k$ along the $w_1 + w_2 + w_3 = 1, 0 \le w_i \le 1$ plane can be seen for the two different dynamic modes.

The corresponding probability distribution for each mode, projected onto the $(w_1, w_2)$-plane, was estimated by Parzen window technique, and can be seen in Fig. 11. The densities are well concentrated to the corners $[1,0,0]$ and $[0,0,1]$ with means $\mathbf{w_{0|1}} = [0.83, 0.11, 0.06]$ and $\mathbf{w_{0|2}} = [0.25, 0.1, 0.65]$ defining the expected weights for each predictor mode.

The nominal mode probability density function was set to $N(\mathbf{\frac{1}{3}\frac{1}{3}\frac{1}{3}}, \mathbf{0.1I})$. In Fig. 11 all density functions, including the nominal mode, projected onto the $(w_1, w_2)$-plane, can be seen together.

*2) Ensemble prediction vs individual predictions:* Using the estimated pdf:s and expected weights $\mathbf{w}$ of the identified predictor modes the ensemble machine was run on the test data. An example of the distribution of the weights for the two dynamic modes $A$ and $B$ can be seen in Fig. 12.

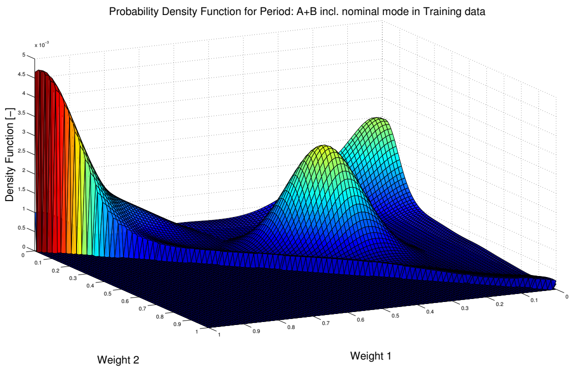

Fig. 11. Estimated probability density functions for the weights in the training data, including nominal mode.

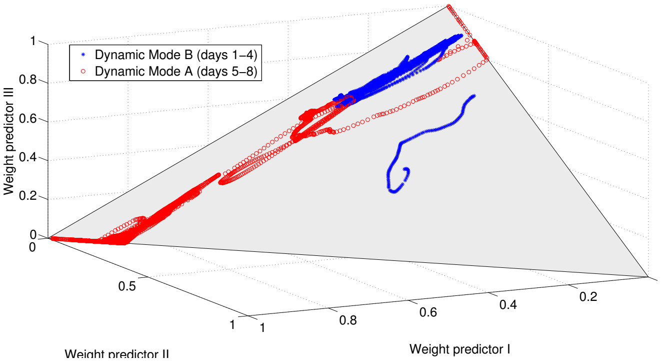

Fig. 12. Example of the distribution of weights in the test data using the estimated pdf:s and expected weights.

An example of how switching between the different modes occurs over the test period can be found in Fig 13.

For evaluation purposes all predictors were run individually. Table V summarizes a comparison of predictive performance between the different approaches in terms of mean Root Mean Square Error (RMSE) over the test batches.

TABLE V
PERFORMANCE EVALUATION FOR THE 60 MINUTE PREDICTORS USING DIFFERENT APPROACHES.

| | RMSE [mg/dl] | | |
|---|---|---|---|
| Predictor Type | Section A | Section B | A+B |
| Predictor I | 8.3 | 16.3 | 13.0 |
| Predictor II | 9.1 | 11.2 | 10.9 |
| Predictor III | 14.3 | 7.9 | 12.6 |
| Merged prediction | 8.7 | 10.5 | 9.6 |

## VI. Comparison to other merging techniques

Compared to the strategy of pure switching between different predictors the evaluation indicates that the proposed algorithm is more robust to sudden changes and in reducing the impact of modeling errors. Apart from that, in many

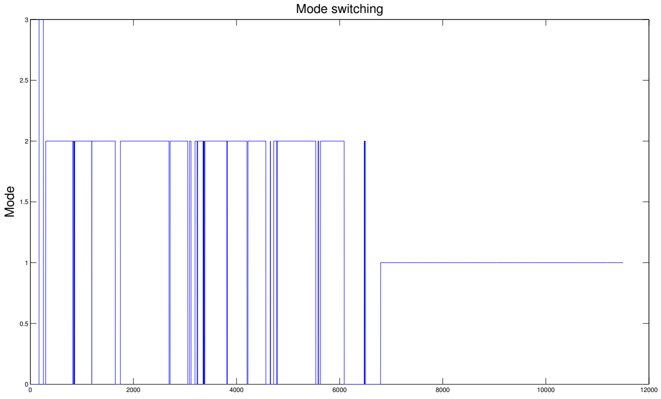

Fig. 13. Example of switching between different predictor modes in the test data. The transition from dynamic mode $B$ to mode $A$ takes place at 5760 min (4 days). The gaps correspond to time instances when no $p_i(w|D)$ fulfilled the criteria.

applications, transition between different dynamic modes is a gradual process rather than an abrupt switch, making the pure switching assumption inappropriate. The proposed algorithm can handle such smooth transitions by slowly sliding along a trajectory in the weight plane of the different predictors, perhaps with a weaker $\Lambda$ if such properties are expected. Furthermore, any type of predictor may be used, not restricting the user to a priori assumptions of the underlying process structure.

In Tagaki-Sugeno (TS) system, a technique that also gives soft switching, the underlying assumption is that the switching dynamics can be observed directly from the data. This assumption has been relaxed for the proposed algorithm extending the applicability beyond that of TS systems.

In [19] another interesting approach to online Bayesian Model Averaging is suggested for changing dynamics. In this approach the assumed transition dynamics between the different modes is based on a Markov chain. However, in our approach no such assumptions on the underlying switching dynamics are postulated. Instead switching is based on recent performance in regards to the applicable norm, and possibly on estimated correlations between predictor modes and features of the data stream $P(\Theta_i|D_k)$, see Eq. 13.

## VII. Conclusions

In this paper a novel merging mechanisms for multiple predictor has been proposed for time-varying conditions. The approach has been evaluated on two different examples of artificial data sets incorporating modeling errors in the individual predictors and different cost criteria. In the first example it was shown how the merged predictor could reduce the impact of the modeling errors resulting in performance beyond the optimal switching strategy. The latter example outlined how the technique may be applied to the specific example of diabetes glucose prediction under sudden changes in the underlying physiological dynamics. Also in this example the merged prediction turned out to be the best choice.

This early assessment indicates that the concept may prove useful when dealing with several individual predictors of uncertain reliability and when data incorporates significant time-varying properties.

Further research will be undertaken to investigate the properties under slowly time-varying processes, the implications of applying different cost functions, and most importantly; performance on real-world data.

## VIII. Acknowledgements

This work has been financially supported by the European FP7 IP IST-216592 DIAdvisor project. Furthermore, the authors are grateful to the Bioengineering Research Group, Dept. of Information Engineering, Padova University, Italy for sharing parameter values for the diabetes simulation model.

## References

[1] Medtronic Diabetes. http://www.medtronicdiabetes.com/products/continuousglucosemonitoring.

[2] A. Alessandri, M. Baglietto, and G. Battistelli. Receding-horizon estimation for switching discrete-time linear systems. *Automatic Control, IEEE Trans. on*, 50(11):1736 – 1748, Nov. 2005.

[3] J. Arenas-Garcia, M. Martinez-Ramòn, A. Navia-Vazquez, and A. R. Figueiras-Vidal. Plant identification via adaptive combination of transversal filters. *Signal Processing*, 86(9):2430 – 2438, 2006. Special Section: Signal Processing in UWB Communications.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, Secaucus, NJ, USA, 2006.

[5] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[6] W. L. Clarke, D.l Cox, L. A. Gonder-Frederick, W. Carter, and S. L. Pohl. Evaluating clinical accuracy of systems for self-monitoring of blood glucose. *Diabetes Care*, 10:622–628, 1987.

[7] C. Dalla Man, R. A. Rizza, and C. Cobelli. Meal simulation model of the glucose-insulin system. *IEEE Trans. Biomed. Eng*, 54(10):1740–1749, 2007.

[8] G. Elliott, C. W.J. Granger, and A. Timmermann, editors. *Handbook of Economic Forecasting*, chapter 10. Forecast Combinations. Elsevier, 2006.

[9] E. J. Elton, M. J. Gruber, and M. W. Padberg. Simple criteria for optimal portfolio selection. *J. of Finance*, 31(5):1341–1357, Dec. 1976.

[10] F. Gustafsson. *Adaptive Filtering and Change Detection*. John Wiley & Sons, Hoboken, New Jersey, USA, 2000.

[11] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averging: A tutorial. *Statistical Science*, 14(4):382–417, 1999.

[12] R. Johansson. *System Modeling & Identification*. KFS AB, 2009.

[13] J. Z. Kolter and M. A. Maloof. Dynamic weighted majority: A new ensemble method for tracking concept drift. *Data Mining, IEEE Int. Conf. on*, 0:123–130, 2003.

[14] B.P. Kovatchev, M. Straume, D.J. Cox, and L.S. Farhy. Risk analysis of blood glucose data: A quantitative approach to optimizing the control of insulin dependent diabetes. *J. of Theor. Med*, 3:1–10, 2000.

[15] Littlestone N. and Warmuth M. K. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, feb. 1994.

[16] H Ohlsson, L. Ljung, and S. Boyd. Segmentation of ARX-models using sum-of-norms regularization. *Automatica*, 46(6):1107–1111, 2010.

[17] N.C. Oza. Online bagging and boosting. In *Systems, Man and Cybernetics, 2005 IEEE Int. Conf. on*, volume 3, pages 2340 – 2345 Vol. 3, Oct. 2005.

[18] A. E. Raftery, T. Gneiting, F. Balabdaoui, and M.L. Pololakowski. Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133:1155–1174, May 2005.

[19] A. E. Raftery, M. Kárný, and P. Ettler. Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics*, 52(1):52–66, Feb 2010.

[20] T. Takagi and M. Sugeno. Fuzzy identification of system and its applications to modelling and control. *IEEE Trans. on Systems, Man, and Cybernetics.*, SMC-15:116–132, Jan-Feb 1985.