# Bayesian averaging of regressive models

**Kamil Dedecius, Ladislav Jirsa, Miroslav Pištěk**

Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 4
CZ-182 08 Prague, Czech Republic
{dedecius,jirsa,pistek}@utia.cas.cz

**Abstract**

In the real world, it is often possible to model certain variables using several different regressive models. However, as the theoretical (mathematical, physical...) description of the real world is almost never perfect, there exists uncertainty about the true or best-fitting model. This paper deals with an issue of 'mixing' information from multiple potentially true models, which run in parallel, to obtain a single outcome, taking the model uncertainty into account. While this issue has been addressed by many research papers in the past, most of them were developed for the static cases or for the state-space models. Here, we discuss an enhancement for a class of input-output regressive models. The described method allows to switch among models to reflect their modelling performance. If there is a single best model, the method quickly converges to it.

## I. Introduction

Standard statistical practice ignores model uncertainty [1]. In a typical modelling scenario, the user selects one model from a whole class of potentially suitable models, preferably using a pre-selected criterion, e.g., the Bayesian information criterion (BIC) or the Akaike's information criterion (AIC), or builds his/her model on the underlying physical theory etc. However, the reality may change and make the model less or more worse than other models in the class. In this case, a technique allowing to switch among multiple models chosen from the class would help to preserve the modelling and prediction abilities.

A suitable method for 'mixing' information from multiple models, called Bayesian model averaging (BMA), was developed in the past, e.g. in [1, 2, 3]. It has been shown, that averaging over multiple or all available models leads to better average predictive ability than any single model from the class. However, the main limitation of the BMA lies in its orientation on static cases. An extension to online dynamic problems was proposed by [4] by means of finite window modelling and in [5] in the form of recursive updating with exponential forgetting [6, 7]. However, in [5], it was derived for the state-space models as dynamic model averaging (DMA). In this paper we discuss a straightforward extension of DMA to regressive models.

Specific notation: $M^{(i)}, i = 1, \ldots, K$ denotes $i$-th model from a set of suitable models $\mathcal{M}$, upper index $(i)$ denotes a quantity or variable related to it. $f(\cdot)$ is probability density function (pdf) determined by its argument with respect to either Lebesgue or counting measure, depending on continuity or discontinuity of the related random variable. $\boldsymbol{Y}^t$ is the data vector, containing the data from the beginning until time $t$. $y_t \in \mathbb{R}$ is a value of interest. $\mathbb{E}[X]$ denotes expected value of argument $X$, $\propto$ is proportionality, i.e. equality up to a constant factor. Time $t = 0, 1, \ldots$ is discrete.

## II. Bayesian model averaging (BMA)

Let's first describe the static case, when we gathered a batch of data until time $t-1$, and the goal is to predict the next value $y_t$[1]. Suppose, that we have $K$ regressive models $M^{(i)}$ from a set of potentially suitable models $\mathcal{M}$. The standard BMA addresses the situation when one of these models, say $M^l$ is correct and its parameter $\boldsymbol{\theta}^{(l)} \in \mathbb{R}^n$ is fixed, but unknown. Let us show, how to combine the information all the selected models carry, to obtain the best available description of the reality.

Let's describe the combined posterior probability density function (pdf) of $y_t$, given past data $\boldsymbol{Y}^{t-1}$, as

$$f(y_t|\boldsymbol{Y}^{t-1}) = \sum_{i=1}^{K} f(y_t|\boldsymbol{Y}^{t-1}, M^{(i)}) f(M^{(i)}|\boldsymbol{Y}^{t-1}) \tag{1}$$

i.e., the average of the posterior pdfs under each of the models considered, weighted by their posterior model probabilities, which, in turn, are given by the relation

$$f(M^{(i)}|\boldsymbol{Y}^{t-1}) \propto f(\boldsymbol{Y}^{t-1}|M^{(i)}) f(M^{(i)}) \tag{2}$$

where

$$f(\boldsymbol{Y}^{t-1}|M^{(i)}) = \int f(\boldsymbol{Y}^{t-1}|\boldsymbol{\theta}^{(i)}, M^{(i)}) f(\boldsymbol{\theta}^{(i)}|M^{(i)}) d\boldsymbol{\theta}^{(i)} \tag{3}$$

is integrated likelihood of model $M^{(i)}$; $\boldsymbol{\theta}^{(i)}$ is a vector of its regression coefficients, $f(\boldsymbol{\theta}^{(i)}|M^{(i)})$ is a prior pdf of parameters $\boldsymbol{\theta}^{(i)}$ under $M^{(i)}$ and $f(M^{(i)})$ is prior probability that $M^{(i)}$ is the true model under the assumption that one of the models from $\mathcal{M}$ is true. All probabilities are implicitly conditional on the set $\mathcal{M}$ of all models being considered [2].

From nonnegativity of the Kullback-Leibler divergence [8] follows, that averaging over all the models provides better predictive ability, as measured by a logarithmic score rule, then any single model $M^{(i)}$ [2, 1]:

$$-\mathbb{E}\left[\ln\left\{\sum_{i=1}^{K} f(y_t|\boldsymbol{Y}^{t-1}, M^{(i)}) f(M^{(i)}|\boldsymbol{Y}^{t-1})\right\}\right] \leq -\mathbb{E}\left[\ln\left\{f(y_t|\boldsymbol{Y}^{t-1}, M^{(i)})\right\}\right] \tag{4}$$

The logarithmic score was suggested by [9]. It assigns each event $A$ a score of $-\ln f(A)$.
The posterior pdfs average (1) has the following combined mean value and variance:

**Mean value**

$$\mathbb{E}\left[y_t|\boldsymbol{Y}^{t-1}\right] \approx \sum_{i=1}^{K} \hat{y}_i f(M^{(i)}|\boldsymbol{Y}^{t-1}), \tag{5}$$

**Variance**

$$\mathrm{var}[y_t|\boldsymbol{Y}^{t-1}] \approx \sum_{i=1}^{K} \left(\mathrm{var}\left[y_t|\boldsymbol{Y}^{t-1}, M^{(i)}\right] + \hat{y}_i^2\right) f(M^{(i)}|\boldsymbol{Y}^{t-1}) - \mathbb{E}\left[y_t|\boldsymbol{Y}^{t-1}\right]^2, \tag{6}$$

where $\hat{y}_i = \mathbb{E}\left[y_t|\boldsymbol{Y}^{t-1}, M^{(i)}\right]$ [10]. This is the keystone of the BMA, however, we need to expand it for online recursive updating.

[1] The use of BMA is not limited only to predictions, of course.

## III. Online approach

The static method is not directly applicable to online modelling of the quantity of interest, however, its expansion to enable it is almost straightforward. For the state-space models, it is derived, e.g., in [5]. This paper describes a generalized approach, suitable for the regressive models in the form

$$f(y_t|\boldsymbol{Y}^{t-1},\boldsymbol{\theta}), \qquad \boldsymbol{\theta}\in\mathbb{R}^n, \tag{7}$$

where the potentially multivariate parameter $\boldsymbol{\theta}$ can be either constant or time-varying. Its estimation is not a part of the topic addressed by this paper; it can be found, e.g., in [6, 11, 12] and many others.

For online decision of the probability of each model, let us introduce the weight of $i$-th model as follows:

$$\omega^{(i)} = f\left(M_{t-1}^{TRUE} = M^{(i)}\middle|\boldsymbol{Y}^{t-1}\right), \qquad \sum_{i=1}^{K}\omega^{(i)} = 1, \qquad \omega^{(i)}\in[0,1], \tag{8}$$

where $M_{t-1}^{TRUE}$ is the true model at time $t-1$.

**Remark 1.** *For better understanding, we will denote the weights with extended time indices in the form $\omega_{t|t}$, where the first index is related to time update and the second to data update of the quantity.*

The initial weights in time instant $t=0$ must be specified by a user. It is possible to use either an expert information about the properties and suitability of each model and set them according to this information (or belief), or to set the weights as uniformly distributed, i.e.

$$\omega_{0|0}^{(i)} = \frac{1}{K}, \qquad i=1,\dots,K \tag{9}$$

As the modelled reality might change, the true model is likely to change as well, which must be reflected by the weights. At this point, we are confronted with two issues:

- time update – time evolution of weights, related to possible change of the true model,
- data update – evolution of the weights caused by the new incoming data.

**Time update** The time update is sometimes referred to as model prediction. It allows us to reflect the transitions between models. The most straightforward approach would be to identify a Markov model of the transitions. In this case, let $\boldsymbol{Q}\in\mathbb{R}^{K\times K}$ be the transition matrix of the model with elements $q_{ij}\in[0,1]$, denoting the probability of transition from model $i$ to model $j$. Then, the rule for prediction of the future true model would read

$$\omega_{t|t-1}^{(i)} = f\left[M_t^{TRUE} = M^{(i)}|\boldsymbol{Y}^{t-1}\right] = \sum_{j=1}^{K}\omega_{t-1|t-1}^{(j)}q_{ij}, \tag{10}$$

However, in most cases the Markov model is unknown and we have to replace (10) with some suitable alternative method. One of them is to let the data speak for themselves [6], increase the uncertainty of $M_t^{TRUE}$ and only modify the weights to prevent their degradation to a singular case, when just one model is preferred over all others. For this purpose, two popular methods were selected – the stabilized exponential forgetting [11, 13] and the linear forgetting [13]. Both these methods modify the distribution of models weights and use an alternative information $f^A(M^{(i)}|\boldsymbol{Y}^{t-1})$ about them. This information can be obtained, e.g., from an expert (cf. initial weights issue discussed above). If there is no useful alternative available, the user can get by just with a flat prior pdf, noninformative pdf etc. Let $\alpha\in[0,1]$ be the forgetting factor; the time update methods are

- stabilized exponential forgetting

$$\omega_{t|t-1}^{(i)} = \frac{\left(\omega_{t-1|t-1}^{(i)}\right)^{\alpha} \left[f^A(M^{(i)}|\boldsymbol{Y}^{t-1})\right]^{(1-\alpha)}}{\sum_{i=1}^{K} \left(\omega_{t-1|t-1}^{(i)}\right)^{\alpha} \left[f^A(M^{(i)}|\boldsymbol{Y}^{t-1})\right]^{(1-\alpha)}} \tag{11}$$

- linear forgetting

$$\omega_{t|t-1}^{(i)} = \frac{\alpha\omega_{t-1|t-1}^{(i)} + (1-\alpha) f^A(M^{(i)}|\boldsymbol{Y}^{t-1})}{\sum_{i=1}^{K} \left[\alpha\omega_{t-1|t-1}^{(i)} + (1-\alpha) f^A(M^{(i)}|\boldsymbol{Y}^{t-1})\right]} \tag{12}$$

Usually, the forgetting factor $\alpha$ is not lower than 0.95.

**Data update** The time update is followed by the data update, incorporating the knowledge about the new obtained data into the weight of each model $M^{(i)}$. This leads to the increase in the weights of the better models to the worse models. We follow the same rule as given in [4]

$$\omega_{t|t}^{(i)} \propto \omega_{t|t-1}^{(i)} f(y_t|\boldsymbol{Y}^{t-1}, M^{(i)}), \tag{13}$$

i.e., the weight of each model is scaled by the model's likelihood and the weights are normalized to sum to one. If the data update step were not accompanied by the time update (forgetting), one model's weight would grow nearly to one, while the others' weights would approach zero. A recovery from such setting would then require a long time run and the modelling would fail.

**Outcome prediction** When the averaged outcome prediction comes into question, the prediction of each model $M^{(i)}, i = 1, \ldots, K$ is calculated and their convex combination is computed. As the coefficients serve the models' weights.

$$\hat{y}_t = \sum_{i=1}^{K} \omega_{t|t}^{(i)}\ \hat{y}_t^{(i)}. \tag{14}$$

## IV. Algorithm

The dynamic Bayesian model averaging of regressive models can be characterized with the following algorithm:

- Initialization

  1. Specify models $M^{(i)}, i = 1, \ldots, K$ where $M^{(i)}$ has the form (7).
  2. Specify initial probabilities of models $\omega_{0|0}^{(i)}$, e.g., uniform – Eq. (9).
  3. Choose a suitable method for time update and, if applicable, a source of an alternative information.

- Online steps, for $t = 1, 2, \ldots$

  1. Measure the regressor values for each model $M^{(i)} \in \mathcal{M}$
  2. Perform parameter filtration and calculate the point estimates of $\hat{\boldsymbol{\theta}}_t^{(i)}; i = 1, \ldots, K$.

3. Predict output of each model $\hat{y}_t^{(i)}$.

$$\hat{y}_t = \sum_{i=1}^{K} \omega_{t|t}^{(i)} \, \hat{y}_t^{(i)} \tag{15}$$

4. Perform the time update of weights of models $\omega_{t|t-1}^{(i)}$ – Eq. (11) or (12).
5. Perform the data update of weights of models $\omega_{t|t}^{(i)}$ – Eq. (13).

## V. Experimental verification

The experimental design is the following – we try to model a mixed time series of 300 samples, created as follows:

$$y_t = 0.8x_t^{(1)} - 0.2 \qquad \text{for } t = (1, \ldots, 100) \cup (201, \ldots, 300), \tag{16}$$

$$y_t = 0.99x_t^{(2)} + 0.5 \qquad \text{for } t = 101, \ldots, 200. \tag{17}$$

where $x^{(1)}$ and $x^{(2)}$ are two different random series generated as random walks with zero mean and unit variance. The mixed output $y$ was further corrupted by additive Gaussian white noise with zero mean and standard deviation 0.2.

The time series was modelled using two linear Gaussian models

$$M^{(1)} : f(y|x^{(1)}, \theta_t^{(1)}), \tag{18}$$

$$M^{(2)} : f(y|x^{(2)}, \theta_t^{(2)}) \tag{19}$$

where the parameters $\theta_t^{(1)}$ and $\theta_t^{(2)}$ were estimated with exponential forgetting with factor 0.99 [6, 7]. The reason for forgetting is justified by the need to eliminate invalid information acquired from data from that part of the series, where the particular model is not the true one. The models' weights were evolved with linear forgetting; as the source of alternative we used the discrete uniform distribution. To make the situation more real, we fed the models with delayed inputs – the delay was 15 steps.

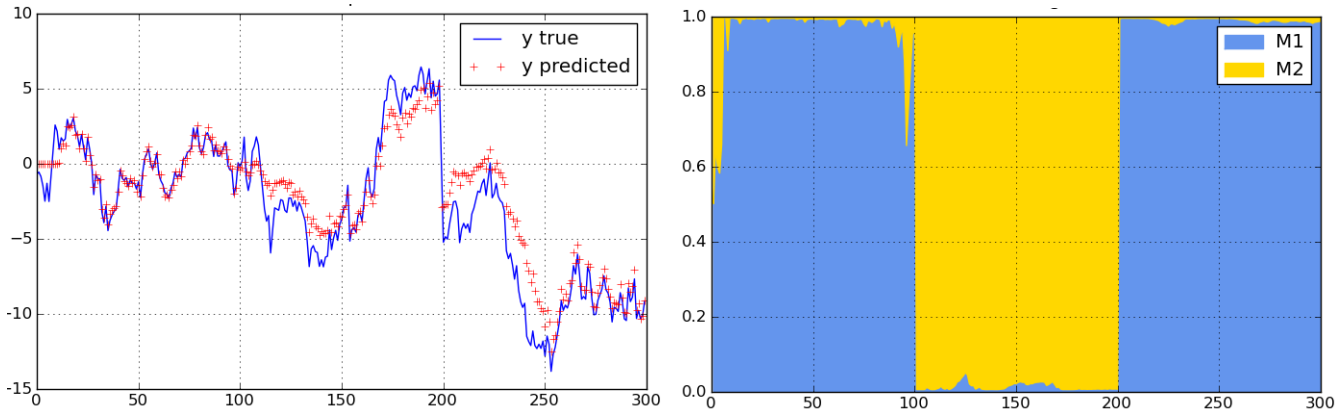
The results are depicted in the Figures 1a and 1b and the statistics of the prediction errors are shown in Table 1. Apparently, the averaged model predicts with better results than any of the two single models itself. Furthermore, a switching between the two models is very fast.

| **Statistics of the error** | **Model 1** | **Model 2** | **Averaged model** |
|---|---|---|---|
| Mean | 0.04 | 2.73 | 0.59 |
| Median | 0.37 | 1.59 | 0.35 |
| Variance | 8.15 | 16.03 | 1.89 |
| Max. negative error | -8.77 | -4.36 | -3.03 |
| Max. positive error | 6.43 | 12.57 | 4.89 |

Table 1: Prediction errors statistics

## VI. Conclusion and future work

A method for online Bayesian model averaging of regression models was discussed. We first introduced the static approach known as the Bayesian model averaging and then derived its recursive variant for input-output regressive models. It allows to mix together outputs of a set of regression models, regardless on the prior knowledge which is the true one at the moment. The impact of the models on




(a) True data and predictions

(b) Evolution of weights

Figure 1: Results of experiment.

prediction is measured by their weights, which quantify the uncertainty related to the models. Besides the setting of the weights according to the models' likelihoods, we employ forgetting to avoid their degradation to a singular case, when just one model dominates the class with a weight close to one. The method will be both theoretically and practically tested in an industrial application – cold sheet rolling mills.

## Acknowledgement

The research was supported by project MŠMT 7D09008 – Probabilistic Bayesian soft sensor - a tool for on-line estimation of the key process variable in cold rolling mills.

## References

[1] A. Raftery, D. Madigan, and J. Hoeting, "Bayesian model averaging for linear regression models," *Journal of the American Statistical Association*, 92(437):179–191, 1997.

[2] J. Hoeting, D. Madigan, A. Raftery, and C. Volinsky, "Bayesian model averaging: A tutorial," *Statistical science*, 14(4):382–401, 1999.

[3] D. Madigan and A. Raftery, "Model selection and accounting for model uncertainty in graphical models using Occam's window," *Journal of the American Statistical Association*, 89(428):1535–1546, 1994.

[4] A. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski, "Using Bayesian model averaging to calibrate forecast ensembles," *Monthly Weather Review*, 133(5):1155–1174, 2005.

[5] A. Raftery, M. Kárný, J. Andrýsek, and Ettler, "Online Prediction under Model Uncertainty Via Dynamic Model Averaging: Application to a Cold Rolling Mill," 2007.

[6] V. Peterka, "Bayesian system identification," *Automatica*, 17(1):41–53, 1981.

[7] A. Jazwinski, *Stochastic processes and filtering theory*, Academic Pr, 1970.

[8] J. Bernardo, "Expected information as expected utility," *The Annals of Statistics*, pp. 686–690, 1979.

[9] I. Good, "Rational decisions," *Journal of the Royal Statistical Society. Series B (Methodological)*, 14(1):107–114, 1952.

[10] A. Raftery, "Bayesian model selection in structural equation models," *Testing structural equation models*, pp. 163–180, 1993.

[11] M. Kárný, *Optimized Bayesian Dynamic Advising: Theory and Algorithms*, Springer-Verlag New York Inc, 2006.

[12] K. Dedecius, I. Nagy, M. Kárný, and L. Pavelková, "Partial Forgetting. A new method for tracking time-variant parameters," 2009.

[13] R. Kulhavý and F. Kraus, "On duality of regularized exponential and linear forgetting," *AUTOMATICA-OXFORD*, 32:1403–1416, 1996.