ADAPTIVE CONTINUOUS HIERARCHICAL MODEL-BASED DECISION MAKING

For Process Modelling with Realistic Requirements

Kamil Dedecius
Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 4, 182 08 Prague, Czech Republic
dedecius@utia.cas.cz

Pavel Ettler
COMPUREG Plzeň, s.r.o., Nádražní 18, 306 34 Plzeň, Czech Republic
ettler@compureg.cz

Keywords: Bayesian modelling, Hierarchical model, Parameter estimation, Cold strip rolling, Rolling mills.

Abstract: Industrial model-based control often relies on parametric models. However, for certain operational conditions either the precise underlying physical model is not available or the lack of relevant or reliable data prevents its use. A popular approach is to employ the black box or grey box models, releasing the theoretical rigor. This leads to several candidate models being at disposal, from which the (often subjectively) prominent one is selected. However, in the presence of model uncertainty, we propose to benefit from a subset of credible models. The idea behind the multimodelling approach is closely related to hierarchical modelling methodology. By using several modelling levels, it is possible to achieve relatively high quality and robust solution, providing a way around typical constraints in industrial applications.

1 INTRODUCTION

Industrial model-based control or prediction is connected with several practical but often contradictory requirements:

- The model should respect physical relations of the process yet be simple enough;
- The model should approximate process behaviour in all possible working conditions;
- Imperfection of measured data feeding the model should not deteriorate control or prediction quality.

Possible solution of the first two above-mentioned issues consists in switching among several proven models according to actual working conditions. The third problem is much more difficult to be solved in principle. The data driven (black box) models could seem to be most appropriate but they almost always contradict the first two requirements. See, e.g., (Bohlin, 1991) for remarks on black and grey box modelling.

On-line mixing of several proven process models – which can be considered as continuous decision making – turns out to be a solid compromise solution of all three requirements, at least for the process of cold strip rolling from which comes the motivation for this research (Ettler and Andrýsek, 2007).

The presented method of model mixing is closely related to dynamic model averaging (Raftery et al., 2010). However, two additional requirements stimulated the need for a specific solution:

- Sum of weights corresponding to particular models is required to be always equal to one;
- An offset term should be continuously estimated to eliminate any residual discrepancy, whatever it comes from.

While the first restriction can be easily justified from the probabilistic point of view, insistence on the second requirement comes purely from practical experience, although the existence of the offset might theoretically be superfluous. Presented solution attempts to reconcile both requirements.

The method is developed in the Bayesian framework, treating the unknown parameters as random variables. Each model is represented by a conditional probability density function (pdf) of the modelled variable, given a set of features and parameters.

The layout of the paper is as follows: Section 2 describes the structure of the multimodel and its levels; in Section 3 we derive a particular case of the model.

Section 4 contains an example of application. Finally, Section 5 concludes the achievements.

2 HIERARCHICAL MODELLING CONCEPT

Assume, that there is a stochastic system observed at discrete time instants $k = 1, 2, \ldots$, which is to be modelled. The statistical framework employs, among others, parametric models, expressing the dependence of the output variable of interest y_k on a nonempty ordered set of observed data

$$\mathcal{D}(k) = \{\boldsymbol{d}_\kappa\}_{\kappa=0,\ldots,k}, \qquad \boldsymbol{d}_\kappa \in \mathbb{R}^n.$$

$\boldsymbol{d}_0$ is the prior knowledge represented, for instance, by an expert information or a noninformative distribution. The task is to predict the future output y_{k+1}, e.g., for control.

Often there is a whole set of available models, mainly those based on underlying physical principles of the process. However, since in many applications either the precise physical model is not available or the lack of reliable data prevents its use, the user employs black box or grey box models. Then, either the (subjectively) best model or models are selected and switched or, alternatively, a rich-structure model being a union of several candidate models is built. Despite the potential applicability of the latter case, the over-parametrization in combination with unreliable measurements and their traffic delays is expected to be fatal.

Our method provides a way around most of disadvantages of the mentioned approaches. We propose a hierarchical model composed of three levels:

Low-level Models comprise arbitrary count of plausible parametric models like the regressive and the space-state ones. They are independent of each other, but their aims are identical – modelling of the same quantity of interest.

Averaging Model is intended for merging the information from the low-level models. The resulting mixture of predictive pdfs of low-level models, weighted by their evidences, is used to evaluate the predictions.

High-level Model – since industry has several specific requirements related, e.g., to stable control, we add a high-level model. It provides stabilization of the prediction process. However, the goal of this level can differ from case to case according to specific needs of the field of application being addressed.

The ensuing sections describe these levels in some detail.

2.1 Low-level Models

The low-level models express the relation between the actual system output y_k and the given data $\mathcal{D}(k)$ by a pdf

$$f(y_k|\mathcal{D}(k-1), \boldsymbol{\Theta}), \tag{1}$$

where $\boldsymbol{\Theta}$ denotes a multivariate finite model parameter which, under the Bayesian treatment, is considered to be a random variable obeying pdf

$$g(\boldsymbol{\Theta}|\mathcal{D}(k-1)). \tag{2}$$

If this pdf is properly chosen from a class conjugate to the model (1), the Bayes' theorem yields a posterior pdf of the same type (Bernardo and Smith, 2001). Then, the rule for recursive incorporation of new measurements into the parameter pdf reads

$$g(\boldsymbol{\Theta}|\mathcal{D}(k)) = \frac{f(y_k|\mathcal{D}(k-1), \boldsymbol{\Theta}) g(\boldsymbol{\Theta}|\mathcal{D}(k-1))}{I_k}, \tag{3}$$

where

$$I_k = \int f(y_k|\mathcal{D}(k-1), \boldsymbol{\Theta}) g(\boldsymbol{\Theta}|\mathcal{D}(k-1)) \mathrm{d}\boldsymbol{\Theta} \tag{4}$$

$$= f(y_k|\mathcal{D}(k-1)) \tag{5}$$

is a normalizing term. It assures unity of the resulting pdf and it is a suitable measure of model's fit, often called *evidence*. The equality of (4) and (5) follows from the Chapman-Kolmogorov equation (Karush, 1961). Furthermore, this equation also yields the predictive pdf $f(y_{k+1}|\mathcal{D}(k))$ providing the Bayesian prediction, formally

$$f(y_{k+1}|\mathcal{D}(k)) = \int f(y_{k+1}|\mathcal{D}(k), \boldsymbol{\Theta}) g(\boldsymbol{\Theta}|\mathcal{D}(k)) \mathrm{d}\boldsymbol{\Theta}$$

$$= \frac{I_{k+1}}{I_k}. \tag{6}$$

The last equality follows from the recursive property of the Bayesian updating (3).

Although the described methodology is important *per se*, it strongly relies on invariance of $\boldsymbol{\Theta}$. However, this assumption is often violated in practical situations and the evolution of $\boldsymbol{\Theta}$ must be appropriately reflected by an additional time update according to model

$$g(\boldsymbol{\Theta}_{k+1}|\boldsymbol{\Theta}_k, \mathcal{D}(k)). \tag{7}$$

Generally, we can distinguish two significant cases:

(i) The evolution model (7) is known *a priori*. Then, $\boldsymbol{\Theta}$ is called the state variable and, under certain conditions, the modelling turns into the famous Kalman filter (Peterka, 1981).

(ii) The model (7) is unknown, but slow variability of Θ is assumed. This case is usually solved either by finite window methods or by forgetting. The latter heuristically circumvents the model ignorance by discounting of potentially outdated information from the parameter pdf. Formally, it introduces a forgetting operator $\mathcal{F}$ modifying the posterior pdf,

$$g(\Theta_{k+1}|\mathcal{D}(k)) = \mathcal{F}\left[g(\Theta_k|\mathcal{D}(k))\right].$$

The class of available forgetting methods comprises, e.g., exponential forgetting (Peterka, 1981), directional forgetting (Kulhavý and Kárný, 1984), partial forgetting (Dedecius, 2010) and others.

Since the issue of parameter variability is behind the scope of the paper, we can stick with unsubscripted Θ without a loss of generality.

2.2 Averaging Model

Assume that there is a nonempty finite set of different low-level models $\mathcal{M} = \{M^{(1)},\dots,M^{(S)}\}$, $S \in \mathbb{N}$, which are considered as candidates to represent the system under study. Formally, we have

$$M^{(s)} : f(y_k|\mathcal{D}(k-1),\Theta^{(s)},M^{(s)}),\ s = 1,\dots,S, \quad (8)$$

which directly coincide with (1). These models are independently evaluated in accordance with Section 2.1. Their probabilities are expressed by a distribution

$$h(\mathcal{M}\,|\mathcal{D}(k)) \equiv h(M^{(1)},\dots,M^{(S)}|\mathcal{D}(k)). \quad (9)$$

The averaging model evaluates this distribution with respect to evidences (5) of low-level models (8) on base of marginal pdfs, namely

$$h(M^{(s)}|\mathcal{D}(k)) \propto h(M^{(s)}|\mathcal{D}(k-1))\ I_k^{(s)}, \quad (10)$$

where $\propto$ denotes equality up to a normalizing factor. The prior distribution $h(M^{(s)}|\mathcal{D}(0))$ can be chosen either on base of expert information or as noninformative pdf with equal marginals.

The predictive pdf of the system output given the set of data $\mathcal{D}(k)$ and the set of models $\mathcal{M}$ is represented by a mixture

$$f(y_{k+1}|\mathcal{D}(k),\mathcal{M}) = \sum_{k=1}^{K} f(y_{k+1}|\mathcal{D}(k),M^{(s)})h(M^{(s)}|\mathcal{D}(k)). \quad (11)$$

The point estimate of y_{k+1} provides the mixture (11) in the form of a convex combination of weighted point estimates $\mathbb{E}\left[y_{k+1}|\mathcal{D}(k),M^{(s)}\right]$. It coincides with the method of Dynamic model averaging (Raftery et al., 2010).

2.3 High-level Model

The purpose of the high-level model is stabilization of the prediction, particularly for its further use in control. Inclusion of the third modelling level is justified by practical experience with averaging models which, in contrast to theoretical assumptions, may provide biased results. The most basic high-level model can be represented by a pdf

$$f(\tilde{y}_{k+1}|\mathcal{D}(k),\hat{y}_{k+1},\tilde{\Theta}), \quad (12)$$

where $\tilde{y}_{k+1}$ is recursively modelled given the features – point estimates of $\hat{y}_{k+1}$ obtained from the averaging model. The high-level model is parametrized by a multivariate parameter $\tilde{\Theta}$. The point estimate $\tilde{y}_{k+1}$ is the output of the hierarchical model. This provides the solution to the task stated in Section 2.

3 ELABORATION FOR INDUSTRIAL APPLICATION

This section elaborates the method for a particular but important case of normal regressive models at the low level. The generalization for another cases, e.g., the state-space models like the Kalman filter and its variants is straightforward and the averaging model remains unchanged. For the sake of convenience, the evolution of parameters will not be discussed.

3.1 Low-level Models

We consider a normal linear regressive model with a regressor $\psi_k \in \mathbb{R}^n$ and a vector of regression coefficients θ of the same dimension, i.e.

$$y_k = \psi_k'\theta + e_k, \quad (13)$$

where $e_k \sim \mathcal{N}(0,\sigma^2)$ is the additive normal white noise. The Bayesian framework relates it with (1) through pdf

$$f(y_k|\mathcal{D}(k-1),\Theta) \sim \mathcal{N}(\psi_k'\theta,\sigma^2), \quad (14)$$

where

$$\Theta \equiv \{\theta,\sigma^2\}.$$

An appropriate distribution conjugate to the model (14) is of the normal inverse-gamma type (Murphy, 2007),

$$g(\Theta|\mathcal{D}(k)) \sim \mathcal{N}i\Gamma(V_k,\nu_k), \quad (15)$$

where $V_k \in \mathbb{R}^{N\times N}$ is an extended information matrix, i.e., a symmetric positive definite square matrix of dimension $N = n+1$, and $\nu_k \in \mathbb{R}^+$ is a number of degrees of freedom. The Bayes' theorem (3) updates these two statistics by new data as follows:

$$\boldsymbol{V}_k = \boldsymbol{V}_{k-1} + \begin{pmatrix} y_k \\ \psi_k \end{pmatrix} \begin{pmatrix} y_k \\ \psi_k \end{pmatrix}'$$
$$\nu_k = \nu_{k-1} + 1.$$

It may be proved (Peterka, 1981) that the estimator of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$ is

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_n \end{pmatrix} = \begin{pmatrix} V_{21} \\ \vdots \\ V_{N1} \end{pmatrix}' \begin{pmatrix} V_{22} & \ldots & V_{2N} \\ \vdots & \ddots & \vdots \\ V_{N2} & \ldots & V_{NN} \end{pmatrix}^{-1}.$$

This relation is equivalent to the recursive least squares (Peterka, 1981) and it provides the estimates for prediction with (13). The normalizing integral $\imath_k$ is nontrivial, it can be found, e.g., in (Peterka, 1981; Kárný, 2006).

Since the information matrices are often ill-conditioned and their inversions can lead to significant numerical issues, it is reasonable to evaluate calculations on factorized forms, e.g., the Cholesky, UDU' or SVD ones.

3.2 Averaging Model

The averaging model introduced in Section 2.2 is responsible for merging the information from the underlaying low-level models. For the sake of better readability, let us denote

$$\alpha_k^{(s)} \equiv h(M^{(s)}|\mathcal{D}(k)).$$

The point prediction of the output at this modelling level follows from (11), hence it reads

$$\hat{y}_{k+1} = \sum_{k=1}^{K} \mathbb{E}\left[y_{t+1}|\mathcal{D}(k), M^{(s)}\right] \alpha_k^{(s)}. \qquad (16)$$

From the statistical viewpoint, two cases can occur. The latter one is a generalization of the former one; both are described below.

3.2.1 Two Low-level Models

Assume that we have two models $M^{(1)}$ and $M^{(2)}$ with real nonnegative scalar statistics $a^{(1)}$ and $a^{(2)}$. We want these models to have probabilities α for $M^{(1)}$ and $1-\alpha$ for $M^{(2)}$. Then, the distribution of models can be viewed as the beta distribution (Gupta and Nadarajah, 2004) with pdf

$$f(\alpha|a^{(1)}, a^{(2)}) = \frac{\Gamma(a^{(1)} + a^{(2)})}{\Gamma(a^{(1)})\Gamma(a^{(2)})} \alpha^{a-1}(1-\alpha)^{a^{(2)}-1}$$

where Γ stands for the gamma function. The point estimate of the mean value and the variance are

$$\hat{\alpha} = \mathbb{E}\left[\alpha|a^{(1)}, a^{(2)}\right] = \frac{a^{(1)}}{a^{(1)} + a^{(2)}}, \qquad (17)$$
$$\mathrm{var}(\alpha) = \frac{a^{(1)}a^{(2)}}{(a^{(1)} + a^{(2)})^2(a^{(1)} + a^{(2)} + 1)}.$$

The rule for update of statistics $a^{(1)}$ and $a^{(2)}$ is as follows:

$$a_k^{(1)} = \alpha_{k-1}\imath_k^{(1)}$$
$$a_k^{(2)} = (1-\alpha_{k-1})\imath_k^{(2)} \qquad (18)$$

3.2.2 Multiple Low-level Models

The distribution of probabilities of multiple models $M^{(1)}, \ldots, M^{(S)}$ can be derived as a generalization of the beta pdf. Let $\boldsymbol{a} = (a^{(1)}, \ldots, a^{(S)})$ be a vector of nonnegative real statistics. Furthermore, let us introduce independent identically distributed (i.i.d.) random variables $W^{(s)} \sim \Gamma(a^{(s)}, 1), s = 1, \ldots, S$ and set

$$\mathbf{W} = (W^{(1)}, \ldots, W^{(S)}), \quad T = \sum_{s=1}^{S} W^{(s)},$$
$$\alpha = (\alpha^{(1)}, \ldots, \alpha^{(S)}) \text{ where } \alpha^{(s)} = \frac{W^{(s)}}{T}. \qquad (19)$$

Obviously, (19) imposes constraints $\alpha^{(s)} \in [0,1]$ and $\sum \alpha^{(s)} = 1$. Since the pdf of a gamma distribution for $W^{(s)}$ is

$$f(W^{(s)}|a^{(s)}, 1) = \frac{1}{\Gamma(a^{(s)})} \left[W^{(s)}\right]^{a^{(s)}-1} e^{-W^{(s)}},$$

the pdf for the multivariate $\mathbf{W}$ with i.i.d. elements has the form

$$f(\boldsymbol{W}|\boldsymbol{a}, \mathbf{1}) = \prod_{s=1}^{S} \frac{1}{\Gamma(a^{(s)})} \left[W^{(s)}\right]^{a^{(s)}-1} e^{-W^{(s)}}.$$

Since α's should sum to unity, we need only $(S-1)$-variate vector $\boldsymbol{\alpha} = (\alpha^{(1)}, \ldots, \alpha^{(S-1)})$. The change of variables theorem (Rudin, 2006) provides a way to interchange of $W^{(s)}$ and $\alpha^{(s)}$,

$$f(\boldsymbol{\alpha}|\cdot) = f_{\boldsymbol{W}}(\boldsymbol{\alpha}|\cdot) \det \boldsymbol{J}_{\boldsymbol{W}\to\boldsymbol{\alpha}}.$$

Here, $\boldsymbol{J}_{\boldsymbol{W}\to\boldsymbol{\alpha}}$ denotes the Jacobian matrix containing partial derivatives of the projection and $f_{\boldsymbol{W}}(\boldsymbol{\alpha}|\cdot)$ is originally a function $f(\boldsymbol{W}|\boldsymbol{a}, \mathbf{1})$ with $\boldsymbol{\alpha}$ substituted for $\boldsymbol{W}$. Since $W^{(s)} = T\alpha^{(s)}$ is bijective for $s = 1, \ldots, S-1$ and $W^{(S)} = T(1 - \alpha^{(1)} - \ldots - \alpha^{(S-1)})$, the necessary condition is fulfilled and the theorem may be used. The determinant of the Jacobian

$$\boldsymbol{J}_{\boldsymbol{W}\to\boldsymbol{\alpha}} = \det\left(\frac{\partial \boldsymbol{W}}{\partial \boldsymbol{\alpha}}, \quad \frac{\partial \boldsymbol{W}}{\partial T}\right) = T^{S-1}$$

provides

$$f(\boldsymbol{\alpha}, T) = T^{A-1}e^{-T} \prod_{s=1}^{S} \frac{1}{\Gamma(a^{(s)})} \left[\alpha^{(s)}\right]^{a^{(s)}-1},$$

where $A = \sum_k a^{(s)}$. Integrating T out with the rule

$$\int T^{A-1} e^{-T} dT = \Gamma(A)$$

leads to the pdf of $\boldsymbol{\alpha}$ as follows:

$$f(\boldsymbol{\alpha}|\boldsymbol{a}) = \frac{\Gamma(A)}{\prod_{s=1}^{S} \Gamma(a^{(s)})} \prod_{s=1}^{S} \left[\alpha^{(s)}\right]^{a^{(s)}-1}$$

This pdf is a variant of the Dirichlet distribution (Geiger and Heckerman, 1997), with nonnegative real statistics $a^{(s)}$. Since the marginal distributions are of beta type $B(a^{(s)}, A-a^{(s)})$, the estimator of kth element of $\boldsymbol{\alpha}$ is

$$\mathbb{E}\left[\alpha^{(s)}\right] = \hat{\alpha}^{(s)} = \frac{a^{(s)}}{A} = \frac{a^{(s)}}{\sum_{s=1}^{S} a^{(s)}}$$

and its variance is

$$\mathrm{var}(\alpha^{(s)}) = \alpha^{(s)} \frac{A - \alpha^{(s)}}{A^2(A+1)}.$$

The rule updating statistics $a^{(s)}$ is a straightforward generalization of (18)

$$a_k^{(s)} = \alpha_{k-1}^{(s)} I_k^{(s)}.$$

3.3 High-level model

The high-level model discussed in Section 2.3 stabilizes the prediction according to specific needs of the application field of interest. In our case, we use a normal regressive model equivalent to (14) with a regressor being composed of the averaging model output $\hat{y}_{k+1}$ (16) and an offset term,

$$\psi_k = (\hat{y}_{k+1}, 1)' \quad \left| \mathcal{D}(k), \mathcal{M} \right. .$$

Its evaluation, i.e., parameter estimation and output prediction, follows from Section 3.1.

4 REAL DATA EXAMPLE

Let us demonstrate the presented method on a simplified example. Used data come from a four-high cold rolling mill and the aim consists in reliable instantaneous prediction of the output strip thickness deviation (denoted h_2), measurable only with a significant time delay. The true evolution of the modelled output strip thickness deviation (h_2) is depicted in Fig. 1. Apparently, we can experience modelling difficulties around $k \approx 800$, where the data abruptly changed.

Three simple low-level models were chosen to

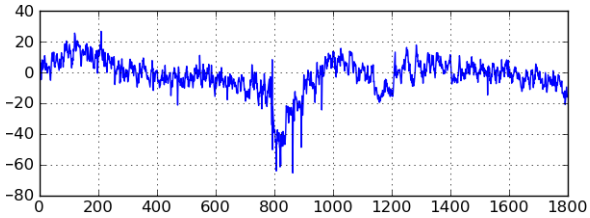

Figure 1: Evolution of the output strip thickness deviation h_2 [μm].

Table 1: Statistics of the prediction error.

Statistics \ Model	averaging	high-level
mean error	0.20	0.04
standard deviation	6.42	5.78
median	0.27	-0.06

approximate relations among selected process variables. They are characterized by the following regressor structures:

$$M^{(1)} : \psi = (v_r, h_1 v_r, 1)' ,$$
$$M^{(2)} : \psi = (h_1, z, 1)' ,$$
$$M^{(3)} : \psi = (h_1, z, v_r, 1)' ,$$

where v_r denotes the ratio of the input and output strip speeds, h_1 is the deviation of the input strip thickness from its nominal value and z stands for the so called uncompensated rolling gap. See (Ettler and Andrýsek, 2007) for details.

The initial setting was as follows: the low-level models started with noninformative prior normal inverse-gamma distributions (15). Forgetting factor of the applied exponential forgetting (Peterka, 1981) was set to 0.99. The averaging model started with uniformly distributed prior statistics $a_0^{(s)}, s = \{1,2,3\}$. The high-level model with the structure given in Section 3.3 started with a noninformative prior pdf with similar initial statistics as the low-level models. Forgetting factor was set to 0.98 in this case.

The evolution of probabilities of the averaged models $M^{(s)}, s = \{1,2,3\}$ is depicted in Fig. 3. The evolution of prediction error for h_2 is depicted in Fig. 2. Obviously, the announced abrupt change in the data course led to higher prediction errors. Statistics of the prediction error stated in Tab. 1 demonstrate the role of the high-level modelling.

5 CONCLUSIONS

A method of multilevel modelling was proposed to improve instantaneous prediction of a key variable in the process of cold strip rolling. The resulting hierarchical model consist of three modelling levels – the

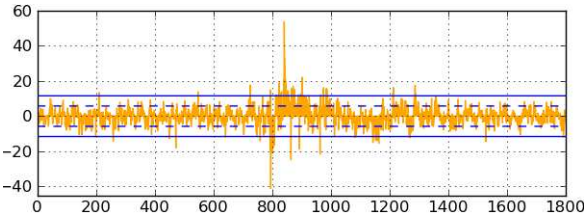

Figure 2: Absolute prediction error [μm]: dashed and solid line denote the distance of one and two standard deviations, respectively.

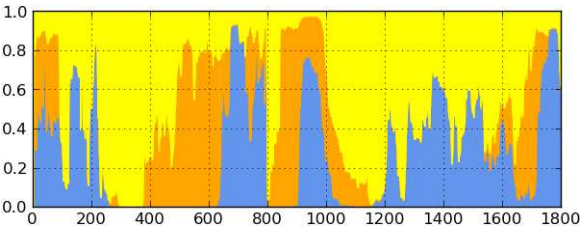

Figure 3: Evolution of probabilities of models: $M^{(1)}$ (blue), $M^{(2)}$ (orange) and $M^{(3)}$ (yellow).

low-level models, i.e., from the statistical viewpoint usual parametric models, the averaging model merging their results and the high-level model, reflecting specific needs of the application field. Each modelling level is treated in the Bayesian framework, i.e., both the models and their parameters are represented by conditional distributions. Current state of the research was demonstrated on a simple example utilizing real industrial data. Extensive tests will be accomplished to refine the method and prepare algorithms for a true on-line industrial application.

ACKNOWLEDGEMENTS

This research was supported by project MŠMT 7D09008 (ProBaSensor) and project 1M0572 (DAR).

REFERENCES

Bernardo, J. and Smith, A. (2001). Bayesian Theory. *Measurement Science and Technology*, 12:221.

Bohlin, T. (1991). *Interactive System Identification: Prospects and Pitfalls*. Springer-Verlag, Berlin, Heidelberg, New York.

Dedecius, K. (2010). *Partial Forgetting in Bayesian Estimation*. PhD thesis, Czech Technical University in Prague.

Ettler, P. and Andrýsek, J. (2007). Mixing Models to Improve Gauge Prediction for Cold Rolling Mills. In *Proceedings of the 12th IFAC Symposium on Automation in Mining, Mineral and Metal Processing*, Quebec, Canada.

Geiger, D. and Heckerman, D. (1997). A Characterization of the Dirichlet Distribution Through Global and Local Parameter Independence. *The Annals of Statistics*, 25(3):1344–1369.

Gupta, A. and Nadarajah, S. (2004). *Handbook of Beta Distribution and its Applications*. CRC.

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

Karush, J. (1961). On the Chapman-Kolmogorov Equation. *The Annals of Mathematical Statistics*, 32(4):1333–1337.

Kulhavý, R. and Kárný, M. (1984). Tracking of Slowly Varying Parameters by Directional Forgetting. In *9th IFAC World Congress, Budapest, Hungary*.

Murphy, K. (2007). Conjugate Bayesian Analysis of the Gaussian Distribution.

Peterka, V. (1981). Bayesian Approach to System Identification In P. Eykhoff (Ed.) Trends and Progress in System Identification.

Raftery, A., Kárný, M., and Ettler, P. (2010). Online Prediction Under Model Uncertainty via Dynamic Model Averaging: Application to a Cold Rolling Mill. *Technometrics: a journal of statistics for the physical, chemical, and engineering sciences*, 52(1):52.

Rudin, W. (2006). *Real and complex analysis*. Tata McGraw-Hill.