Dynamic Diffusion Estimation in Exponential Family Models

Kamil Dedecius and Vladimíra Sečkárová

***Abstract*—This letter proposes a new dynamic diffusion estimation method for a collaborative inference of a common model parameter using a distributed network of cooperating nodes. Unlike the existing single problem-oriented diffusion methods, it is formulated abstractly for the exponential family of models. The resulting advantage—its easy and straightforward application to the family members—is demonstrated on three selected cases: i) the diffusion autoregression, ii) the diffusion Poisson modelling and iii) the diffusion estimation of a Bernoulli process with unknown proportions. The first case is shown to coincide with the diffusion recursive least squares.**

***Index Terms*—Diffusion estimation, distributed estimation, parameter estimation, sensor networks.**

I. Introduction

We address the dynamic distributed estimation of an unknown parameter of interest from noisy measurements by a diffusion network. Each node exchanges information on observations and estimates with its adjacent neighbors and incorporates it locally into its own statistical knowledge. This significantly improves the statistical properties and robustness of the estimation process under regular conditions [1]. Unlike the consensus algorithms and their variations, e.g. [2]–[5], the diffusion algorithms do not require multiple intermediate iterations between two subsequent measurements, see, e.g. [1]. Furthermore, Tu and Sayed [6] show that the diffusion strategies can outperform the consensus strategies in dynamic environments.

The diffusion solutions are mostly least-squares (LS) oriented, for instance the diffusion least mean squares (LMS) [7], [8], recursive least squares (RLS) [1] or the Kalman filter [9]. Although otherwise sound, they are strongly single-problem oriented and their reformulation for other tasks, e.g. non-LS oriented, is limited or even impossible by nature. The goal of this letter is to overcome this shortcoming. By exploiting the consistent theory of the Bayesian inference, we formulate a new dynamic diffusion estimation method in an abstract way, theoretically independent of a particular model type. The only assumption is its membership in the exponential family. Examples are the normal regression models, Poisson (shot noise) model, Bernoulli, Weibull, Pareto and many other models. We note that the dynamic estimation of a varying parameter coincides with the (Bayesian) parameter tracking. Since the proposed distributed estimation method is rooted in this realm, it is directly possible to use most of the elaborated Bayesian tracking methods, for instance forgetting, e.g. [10], [11] and the references therein.

II. Bayesian Estimation in Exponential Family

Consider discrete-time dynamic modelling of an observed variable y_t determined by an unknown fixed parameter θ and, if exists, a known explanatory variable (e.g. regressor) x_t.[1] $t = 1, 2, \ldots$ are time indices. From the probabilistic viewpoint, the model can be represented by a conditional probability density function (pdf) $f(y_t|x_t, \theta)$. Estimation of θ is based on the knowledge of past data $D_{t-1} = \{y_\tau, x_\tau\}_{\tau=1,\ldots,t-1}$ and the prior pdf $\pi(\theta|D_0)$, obtained, e.g., from an expert, based on historical data, or alternatively being a flat noninformative pdf. The Bayesian approach to estimation recursively updates the prior pdf by new data $\{y_t, x_t\}$ via the Bayes' rule [12],

$$\pi(\theta|D_t) \propto f(y_t|x_t, \theta)\pi(\theta|D_{t-1}). \tag{1}$$

Here $\propto$ stands for proportionality, i.e. equality up to a normalizing factor. We call (1) the *sequential* variant. Equivalently, for time horizon t, the *batch* estimation reads

$$\pi(\theta|D_t) \propto \pi(\theta|D_0) \prod_{\tau=1}^{t} f(y_\tau|x_\tau, \theta). \tag{2}$$

Analytical tractability of recursions (1), (2) is guaranteed if the model $f(y_t|x_t, \theta)$ is an exponential family distribution and the prior pdf is conjugate to it, as defined below [12] (with time indices dropped):

Definition 1 (Exponential Family of Distributions): An exponential family of distributions of a variable y with a parameter θ and an explanatory variable x is a family of distributions with pdf of the form

$$f(y|x, \theta) = h(y, x)g(\theta)\exp\left[\eta(\theta)T(y, x)\right], \tag{3}$$

where $h(y, x)$ is a known function, $g(\theta)$ is a known normalization function, $\eta(\theta)$ is a natural parameter and $T(y, x)$ is a sufficient statistic.

Definition 2 (Conjugate Prior pdf): A conjugate prior pdf for a parameter θ with the hyperparameters ξ of the same dimension as $T(y, x)$ and $\nu \in \mathbb{R}^+$ has the form

$$\pi(\theta|\xi, \nu) = q(\xi, \nu)g(\theta)^\nu \exp\left[\eta(\theta)\xi\right], \tag{4}$$

where $q(\xi, \nu)$ is a normalization function and $g(\theta)$ has the same form as in the exponential family.

Manuscript received July 03, 2013; accepted September 01, 2013. Date of publication September 16, 2013; date of current version September 20, 2013. This work was supported in part by the Czech Science Foundation under Grant 13-13502S and by MŠMT under Grant 7D12004. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Alexandros G. Dimakis.

The authors are with the Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Prague 18208, Czech Republic (e-mail: dedecius@utia.cas.cz; seckarov@utia.cas.cz).

Digital Object Identifier 10.1109/LSP.2013.2282042

[1] For the sake of generality, the variables are considered real, possibly multivariate and with compatible dimensions.

The dimension-preserving sufficient statistic accumulates all statistical knowledge necessary to compute an estimate of θ, regardless of the data sample size. It has the form $T(y_t, x_t)$ for the sequential variant (1), and $T(D_t) = \sum_{\tau=1}^{t} T(y_\tau, x_\tau)$ for the product in the batch variant (2). The sequential update modifies the prior hyperparameters as follows:

$$\begin{aligned} \xi_t &= \xi_{t-1} + T(y_t, x_t) \\ \nu_t &= \nu_{t-1} + 1, \end{aligned} \tag{5}$$

similar rules hold for (2). For simplicity, we stick with the sequential variant in the sequel. The modifications for the batch variant are straightforward.

The point estimate of θ can be obtained from the posterior pdf using standard formulas. Usually it is the mean value; sometimes the median or the mode are preferred. The estimation uncertainty is often expressed by the estimator variance.

III. Diffusion Estimation

The diffusion network is an undirected connected graph of N spatially distributed nodes (e.g. sensors). Each node $i = 1, \ldots, N$ can directly exchange information with adjacent nodes forming its closed neighborhood $\mathcal{N}_i$ of a cardinality n_i; also $i \in \mathcal{N}_i$. The exchanged information relates to (i) observations (adapt step) and (ii) estimates (combine step). Since the information from the nodes $j \in \mathcal{N}_i$ may have different credibility from the ith node's viewpoint, nonnegative relative weights summing to unity are used to reflect this.

A. Adapt Step

Each network node $i = 1, \ldots, N$ employs the same form of an exponential family model $f_i(y_t|x_{i,t}, \theta)$ as above. Fixing i and t, we may regard $f_j(y_t|x_{j,t}, \theta)$ for $j \in \mathcal{N}_i$ as a complete system of hypotheses about the true model at i. From the ith node's viewpoint, these are valid with probabilities c_{ij}, called weights, summing to unity due to the completeness. The Kullback-Leibler (KL) divergence [12] defined in Appendix in the role of the loss function then provides the way to approach the true model by pdf f_i^* as follows:

Proposition 1: Given pdfs f_j with weights c_{ij}, $j \in \mathcal{N}_i$, the best approximating pdf f_i^* optimal in the KL sense, minimizing the cumulative loss

$$\sum_{j \in \mathcal{N}_i} c_{ij} \mathrm{D}\left(f_i^* \| f_j\right)$$

has the form

$$f_i^* \propto \prod_{j \in \mathcal{N}_i} f_j^{c_{ij}}.$$

Proof: By definition of the KL divergence

$$\begin{aligned} &\sum_{j \in \mathcal{N}_i} c_{ij} \int_{y_t} f_i^*(y_t|\cdot) \log \frac{f_i^*(y_t|\cdot)}{f_j(y_t|\cdot)} \mathrm{d}y_t \\ &= \int_{y_t} f_i^*(y_t|\cdot) \log \frac{f_i^*(y_t|\cdot)}{\prod_{j \in \mathcal{N}_i} f_j(y_t|\cdot)^{c_{ij}}} \mathrm{d}y_t \\ &= \mathrm{D}\left(f_i^* \middle\| \prod_{j \in \mathcal{N}_i} f_j^{c_{ij}} \right). \end{aligned}$$

The minimum of the KL divergence is attained when its arguments agree. ■

The KL-optimal model f_i^* is given by the geometric mean of available hypothetical models. The initial choice of exponential family models yields the appealing consequence of analytically tractable recursive diffusion update rules similar to (5). The Bayes' theorem (1) with $f = f_i^*$ updates the hyperparameters according to the following proposition.

Proposition 2 (Adapt-Posterior pdf): Given sufficient statistics $T(y_{j,t}, x_{j,t})$, $j \in \mathcal{N}_i$, the adapt step updates the ith node's hyperparameters $\xi_{i,t-1}$ and $\nu_{i,t-1}$ as follows

$$\begin{aligned} \xi_{i,t} &= \xi_{i,t-1} + \sum_{j \in \mathcal{N}_i} c_{ij} T(y_{j,t}, x_{j,t}) \\ \nu_{i,t} &= \nu_{i,t-1} + 1. \end{aligned} \tag{6}$$

The proof is trivial.

Remark 1: The KL divergence is a well founded measure of pdfs' dissimilarity [12]. The chosen zero-forcing order of its arguments brings the salient feature of analytically tractable computations in the exponential family due to the geometric mean, at the potential cost of variance underestimation. The alternative order (zero-avoiding divergence) would yield the arithmetic average of pdfs, raising computational issues and potential variance overestimation [13].

B. Combine Step

The combine step follows the adapt step in order to further improve the statistical properties of individual estimators. We propose two principally different methods, one combining whole adapt-posterior pdfs, the other combining only the point estimates.

a) Whole adapt-posterior pdfs: the ith node combines the adapt-posterior pdfs with the hyperparameters (6) of nodes $j \in \mathcal{N}_i$ in the KL-optimal sense prescribed by Proposition 1 (with the posterior pdfs π_j and weights a_{ij} in the roles of f_j and c_{ij}). The resulting combine-posterior pdf π_i^* has the following hyperparameters,

$$\begin{aligned} \xi_{i,t}^* &= \sum_{j \in \mathcal{N}_i} a_{ij} \xi_{j,t} \\ \nu_{i,t}^* &= \sum_{j \in \mathcal{N}_i} a_{ij} \nu_{j,t}. \end{aligned} \tag{7}$$

These hyperparameters completely characterize the distribution of θ. It is usually very easy to evaluate its moments, quantiles etc. using standard formulas. Furthermore, the combine-posterior pdf can serve as the prior for the next adapt step at $t + 1$.

b) Point estimates: if the ith node has access only to the point estimates provided by nodes $j \in \mathcal{N}_i$, for instance the means $\hat{\theta}_{j,t}$ and optionally the related variances $\gamma_{j,t}$, it is possible to directly combine them as follows:

$$\begin{aligned} \hat{\theta}_{i,t}^* &= \sum_{j \in \mathcal{N}_i} a_{ij} \hat{\theta}_{j,t} \\ \gamma_{i,t}^* &= \sum_{j \in \mathcal{N}_i} a_{ij} \left[\gamma_{j,t} + \left(\hat{\theta}_{j,t} - \hat{\theta}_{i,t}^* \right)^2 \right]. \end{aligned} \tag{8}$$

This approach, motivated by the mixture-based estimation [14], is slightly computationally cheaper, because it avoids intermediate combination of pdfs. The adapt-posterior pdfs remain unmodified and enter $t + 1$ as the prior.

TABLE I
WEIGHTS BEFORE NORMALIZATION. $n_i = \text{cardinality}(\mathcal{N}_i)$. THE SAME WEIGHTS CAN BE USED FOR c_{ij}

Method	**Rule**
Uniform	$a_{ij} = 1/n_i$
Laplacian	$a_{ij} = 1/n_{\max}$
Maximum degree	$a_{ij} = 1/N$
Metropolis	$a_{ij} = 1/\max(n_i, n_j)$
Relative degree	$a_{ij} = n_j / \left(\sum_{k\in\mathcal{N}_i} n_k\right)$
Rel. degree-noise variance	$a_{ij} = n_j\sigma_j^2 / \left(\sum_{k\in\mathcal{N}_i} n_k\sigma_k^2\right)$

C. Choice of Weights

The purpose of weights a_{ij} and c_{ij} is to express the ith node's degree of belief in information from the nodes $j \in \mathcal{N}_i$. For fixed i, both a_{ij} and c_{ij} sum to unity. There exist several (mostly static) methods for their determination, some of them are given in Table I, see, e.g. [15] and references therein. An additional feature of the chosen probabilistic framework is the prospect of theoretically justified information-based methods for dynamic weights. For instance, it is possible to exploit local modelling and sharing of the observations/estimators variances at each node or to measure the fit of the data/estimates using the likelihoods. However, this issue is beyond the main message of the letter and will be addressed in the future.

IV. EXAMPLES

A. Diffusion Autoregression

Consider a Kth order autoregressive model

$$y_t = x_t'\beta = \sum_{k=1}^{K} y_{t-k}\beta_k + \varepsilon_t, \qquad (y_t \in \mathbb{R}^1),$$

where the explanatory variable $x_t = [y_{t-1}, \ldots, y_{t-K}]' \in \mathbb{R}^K$ is a known column regression vector, $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ is additive white noise and $\beta = [\beta_1, \ldots, \beta_K]' \in \mathbb{R}^K$ is a column vector of unknown regression coefficients. Its estimation provides, among others, the least squares (LS) method via the normal equations; the recursive variant is RLS. The same point estimator follows from the Bayesian modelling with $y_t \sim \mathcal{N}(x_t'\beta, \sigma^2)$ and a normal prior distribution for the parameter $\theta \equiv \beta$; the associated uncertainty is an inherent part of the solution. We focus on a bit more complex normal inverse-gamma prior distribution $\mathcal{N}i\mathcal{G}(V, \nu)$, providing an additional advantage of variance estimation with $\theta \equiv \{\beta, \sigma^2\}$. Its hyperparameters standing in the roles of ξ and ν are the extended (symmetric) information matrix $V \in \mathbb{R}^{(K+1)\times(K+1)}$ and the degrees of freedom $\nu \in \mathbb{R}^+$ [12].

Let us demonstrate the ease of derivation of the diffusion estimator. The model pdf in the vector form reads

$$f(y_t|x_t, \theta) = \frac{\sigma^{-1}}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2} \begin{bmatrix} -1 \\ \beta \end{bmatrix}' \begin{bmatrix} y_t \\ x_t \end{bmatrix} \begin{bmatrix} y_t \\ x_t \end{bmatrix}' \begin{bmatrix} -1 \\ \beta \end{bmatrix} \right\}.$$

A rearrangement of the terms according to (3) reveals the sufficient statistic connected with time t,

$$T(y_t, x_t) = \begin{bmatrix} y_t \\ x_t \end{bmatrix} \begin{bmatrix} y_t \\ x_t \end{bmatrix}'. \tag{9}$$

Hence the update (5) of the $\mathcal{N}i\mathcal{G}$ hyperparameters takes the form

$$V_t = V_{t-1} + \begin{bmatrix} y_t \\ x_t \end{bmatrix} \begin{bmatrix} y_t \\ x_t \end{bmatrix}' \qquad \text{and} \qquad \nu_t = \nu_{t-1} + 1.$$

Recall that the autoregressive recursion begins with $t = K + 1$, imposing the initialization with $\nu_K = \nu_0$ and $V_K = V_0$. For $t \geq K + 1$, the point estimators of β and σ^2 are easily reachable after partitioning the matrix V into blocks [10]

$$V_t \equiv \begin{bmatrix} V_{yy,t} & V_{yx,t}' \\ V_{yx,t} & V_{xx,t} \end{bmatrix}, \qquad V_{yy,t} \in \mathbb{R}^1.$$

Then

$$\hat{\beta}_t = V_{xx,t}^{-1} V_{yx,t} \quad \text{and} \quad \hat{\sigma}_t^2 = \frac{V_{yy,t} - V_{yx,t}' V_{xx,t}^{-1} V_{yx,t}}{\nu_t - K + 2}. \tag{10}$$

The diffusion estimator is as follows: The *adapt step* prescribed by Proposition 2 has the form

$$\begin{aligned} V_{i,t} &= V_{i,t-1} + \sum_{j\in\mathcal{N}_i} c_{ij} \begin{bmatrix} y_{j,t} \\ x_{j,t} \end{bmatrix} \begin{bmatrix} y_{j,t} \\ x_{j,t} \end{bmatrix}' \\ \nu_{i,t} &= \nu_{i,t-1} + 1. \end{aligned} \tag{11}$$

The *combine step* is a direct application of the prescribed rules, too. The first case, the *whole adapt-posterior pdfs* combination using (7) and (11) reads

$$\begin{aligned} V_{i,t}^* &= \sum_{j\in\mathcal{N}_i} a_{ij} V_{j,t} \\ \nu_{i,t}^* &= \sum_{j\in\mathcal{N}_i} a_{ij} \nu_{j,t}. \end{aligned}$$

The *point estimates* combination puts (10) into (8).

This diffusion autoregression (with the *point estimates* combine method) coincides with the diffusion RLS proposed by Cattivelli *et al.* [1]. This is proved and discussed in *Supplementary material.* Additionally, it provides the noise variance estimator, which can be potentially useful for dynamic determination of the relative degree-noise variance weights (Table I). The notable benefit of the proposed Bayesian approach over the non-Bayesian one lies in the ease and straightforwardness of its application to a chosen problem while still completely retaining all theoretical consistency.

B. Homogeneous Poisson Process

The homogeneous Poisson process (alias homogeneous shot noise) is a random process $\{M_t\}_{t\geq 0}$ of the counts $M_t \in \mathbb{N}_0$, starting with $M_0 = 0$ and with independent stationary Poisson distributed increments satisfying

$$\mathbb{P}[M_{t+\tau} - M_t = y_t|\lambda] = \frac{(\lambda\tau)^{y_t} e^{-\lambda\tau}}{y_t!}, \qquad \tau \in \mathbb{N}. \tag{12}$$

The *rate* parameter $\lambda \in \mathbb{R}^+$ coincides with the mean and variance of y_t. The process characterizes, e.g., the number of photons or other particles incident on a detector.

Considering the sequential variant with $\tau = 1$ and rewriting (12) to the form (3) reveals the sufficient statistic

$$T(y_t) = y_t.$$

The conjugate prior for $\theta \equiv \lambda$ is the gamma distribution $\mathcal{G}(\alpha, \beta)$ with shaping hyperparameters $\alpha, \beta > 0$ in the roles of ξ and ν, respectively. Their update (5) has the form [12]

$$\begin{aligned} \alpha_t &= \alpha_{t-1} + y_t \\ \beta_t &= \beta_{t-1} + 1. \end{aligned} \tag{13}$$

The point estimator of λ is well known to be $\hat{\lambda}_t = \alpha_t/\beta_t$ with the variance $\gamma_t = \alpha_t/\beta_t^2$.

Now we easily derive the diffusion estimator. The *adapt step* according to Proposition 2 reads

$$\begin{aligned} \alpha_{i,t} &= \alpha_{i,t-1} + \sum_{j \in \mathcal{N}_i} c_{ij} y_{j,t} \\ \beta_{i,t} &= \beta_{i,t-1} + 1. \end{aligned} \tag{14}$$

The *combine step* for *whole adapt-posterior pdfs* (7) reads

$$\begin{aligned} \alpha_{i,t}^{*} &= \sum_{j \in \mathcal{N}_i} a_{ij} \alpha_{j,t} \\ \beta_{i,t}^{*} &= \sum_{j \in \mathcal{N}_i} a_{ij} \beta_{j,t}. \end{aligned} \tag{15}$$

If only the combination of *point estimates* is required, then (8) with the above given point estimators is used.

C. Estimation of Bernoulli Process Proportions

This example studies the Bernoulli process exploited, e.g., in the queuing theory, reliability analysis and finance. It is a discrete-time stochastic process yielding a sequence of independent identically distributed binary random variables Y_t taking values 0 or 1 (failure or success). It follows the Bernoulli distribution,

$$\mathbb{P}(Y_t = y_t | p) = p^{y_t}(1-p)^{1-y_t}, \qquad y_t \in \{0, 1\},$$

where $p \in [0, 1]$ is the probability of success $(y_t = 1)$. Clearly $T(y_t) = y_t$. The conjugate prior for unknown parameter $\theta = p$ is the beta distribution $\mathcal{B}(\alpha, \beta - \alpha)$ with the hyperparameters $\alpha, \beta > 0$ in the roles of ξ and ν, respectively. Their update (5) is

$$\begin{aligned} \alpha_t &= \alpha_{t-1} + y_t \\ \beta_t &= \beta_{t-1} + 1. \end{aligned} \tag{16}$$

The point estimator is known to be $\hat{p}_t = \alpha_t/\beta_t$ with the variance $\gamma_t = \alpha_t(\beta_t - \alpha_t)/[\beta_t^2(\beta_t + 1)]$.

Note the appealing fact arising from the Bayesian estimation of exponential family models with conjugate priors: the recursions (13) and (16) are identical although the underlying distributions are not. The diffusion estimation *adapt* and *combine* steps would accordingly agree with (14) and (15) (or the combination of *point estimates* (8)).

V. Estimators Properties

Generally, the (unique) Bayes estimators are admissible in that there exists no other rule that dominates them with respect to the selected risk function. For instance, under MSE, the Bayes' rule is unique and admissible. Furthermore, it is also asymptotically unbiased, consistent and efficient. More on this can be found, e.g., in [16]. If the nodes provide correct information on the estimated parameter, then the properties hold in the diffusion algorithm as well; otherwise, the result is a shrinkage estimator [16]. A deep study of the diffusion estimator statistical properties is a part of further research.

Appendix
Kullback-Leibler Divergence

Given two pdfs f and g, their Kullback-Leibler (KL) divergence is the nonnegative functional

$$\mathrm{D}(f\|g) = \mathbb{E}_{f(\zeta)}\left[\log \frac{f(\zeta)}{g(\zeta)}\right] = \int_{\zeta} f(\zeta) \log \frac{f(\zeta)}{g(\zeta)} \mathrm{d}\zeta.$$

Properties: $\mathrm{D}(f\|g) = 0$ iff $f = g$ a.e. and $\mathrm{D}(f\|g) \neq \mathrm{D}(g\|f)$ if $f \neq g$. The triangle inequality does not hold.

References

[1] F. S. Cattivelli, C. G. Lopes, and A. H. Sayed, "Diffusion recursive least-squares for distributed estimation over adaptive networks," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1865–1877, May 2008.

[2] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proc. IEEE*, vol. 95, no. 1, pp. 215–233, Jan. 2007.

[3] I. D. Schizas, A. Ribeiro, and G. B. Giannakis, "Consensus in *ad hoc* WSNs with noisy links—Part I: Distributed estimation of deterministic signals," *IEEE Trans. Signal Process.*, vol. 56, no. 1, pp. 350–364, Jan. 2008.

[4] C. Mosquera, R. López-Valcarce, and S. Jayaweera, "Stepsize sequence design for distributed average consensus," *IEEE Signal Process. Lett.*, vol. 17, no. 2, pp. 169–172, 2010.

[5] A. Jadbabaie, P. Molavi, A. Sandroni, and A. Tahbaz-Salehi, "Non-Bayesian social learning," *Games Econ. Beh.*, vol. 76, no. 1, pp. 210–225, Sep. 2012.

[6] S.-Y. Tu and A. H. Sayed, "Diffusion strategies outperform consensus strategies for distributed estimation over adaptive networks," *IEEE Trans. Signal Process.*, vol. 60, no. 12, pp. 6217–6234, Dec. 2012.

[7] C. G. Lopes and A. H. Sayed, "Diffusion least-mean squares over adaptive networks: Formulation and performance analysis," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3122–3136, Jul. 2008.

[8] Y. Liu, C. Li, W. K. S. Tang, and Z. Zhang, "Distributed estimation over complex networks," *Inf. Sci.*, vol. 197, pp. 91–104, Aug. 2012.

[9] F. S. Cattivelli and A. H. Sayed, "Diffusion strategies for distributed Kalman filtering and smoothing," *IEEE Trans. Autom. Control*, vol. 55, no. 9, pp. 2069–2084, Sept. 2010.

[10] V. Peterka, "Bayesian approach to system identification," in *In Trends and Progress in System Identification*, P. Eykhoff, Ed. Oxford, U.K.: Pergamon, 1981, pp. 239–304.

[11] K. Dedecius, I. Nagy, and M. Kárný, "Parameter tracking with partial forgetting method," *Int. J. Adap. Contr. Signal Process.*, vol. 26, no. 1, pp. 1–12, Jan. 2012.

[12] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*, 1st ed. New York, NY, USA: Wiley, 1994.

[13] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. New York, NY, USA: Springer-Verlag, 2006.

[14] S. Frühwirth-Schnatter, *Finite Mixture and Markov Switching Models (Springer Series in Statistics)*, 1st ed. Berlin, Germany: Springer, Aug. 2006.

[15] F. S. Cattivelli and A. H. Sayed, "Diffusion LMS strategies for distributed estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1035–1048, Mar. 2010.

[16] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. Berlin, Germany: Springer, Aug. 1998.