Sequential estimation and diffusion of information over networks: A Bayesian approach with exponential family of distributions

Kamil Dedecius, *Member, IEEE,* and Petar M. Djurić, *Fellow, IEEE*

***Abstract*—Diffusion networks where nodes collaboratively estimate the parameters of stochastic models from shared observations and other estimates have become an established research topic. In this paper the problem of sequential estimation where information in the network diffuses with time is formulated abstractly and independently from any particular model. The objective is to reach a generic solution that is applicable to a wide class of popular models and based on the exponential family of distributions. The adopted Bayesian and information-theoretic paradigms provide probabilistically consistent means for incorporation of shared observations in the implemented estimation of the unknowns by the nodes as well as for effective combination of the "knowledge" of the nodes over the network. It is shown and illustrated on four examples that under certain conditions, the resulting algorithms are analytically tractable, either directly or after simple approximations. The examples include the linear regression, Kalman filtering, logistic regression, and the inference of an inhomogeneous Poisson process. The first two examples have their more or less direct counterparts in the state-of-the-art diffusion literature whereas the latter two are new.**

***Index Terms*—Diffusion network, diffusion estimation, adaptation, combination, exponential family.**

I. Introduction

NETWORKS of interconnected agents collaboratively solving distributed inference problems have attained a considerable research interest in the last decade. This is due to the wide variety of possible applications, including environment monitoring, disaster relief management, source localizations, precision agriculture, and medical applications [1]–[4]. The nodes in the network share with their neighbors their private signals and all the information about the variables of interest, thereby collectively improving the inference results.

The distributed processing schemes for learning over adaptive networks can be classified into three types of strategies: incremental [5]–[11], diffusion, [12]–[17], and consensus-based strategies [18]–[23]. The incremental strategies exploit communication on a cyclic path by connecting all the network nodes in a Hamiltonian cycle. Because each node and link in the network represent a single point of failure, the robustness of this setup is limited and the recovery from a failure requires a redesign of the network, which is an NP-hard problem [3].

The diffusion strategies, on the other hand, rely on networks represented by directed or undirected connected graphs with node degrees usually higher than one. Instead of a cyclic path, the shared information gradually diffuses through the whole network by local communication among adjacent nodes [24]. Typically, the diffusion strategies at each time step are composed of two phases – (a) an *adaptation* that updates the estimates of a node by its neighbors' observations, and (b) a *combination* that merges the neighbors' estimates. One can argue that this strategy imitates the behavior of many self-organized real-world systems, and it is scalable and highly robust to node or link failures. It also preserves the functionality of the network through graceful degradation.

The consensus-based strategies are somewhat similar to the diffusion strategies. The original consensus strategies rely on two time scales: (i) sensing, that is, acquisition of observations across the nodes, and (ii) collaborative processing of the data through iterations, with the goal of achieving a consensus in the computation of a desired, e.g., average value. The diffusion strategies, however, perform both their phases intrinsically on a single time scale, without the need for any intermediate iterative steps. We point out though that recent consensus algorithms remove the requirement for iterations, which brings these methods closer in spirit to the diffusion methods (e.g., the running consensus method [25]). A particular example is the gossip interactive Kalman filter [26], where each node runs a Kalman filter. At random times, a node randomly selects a neighbor and the two swap their states. At any time instant, a node represents a suboptimal local data fusion center, which with time incorporates information in measurements coming from different parts of the network that are randomly selected. This may be viewed as a counterpart of the diffusion combination step, but there is no counterpart of the adaptation step.

Numerous diffusion-based inference algorithms were proposed in the last decade, most of them based on the least-squares principle. The basic diffusion that uses the least mean squares (LMS) criterion was originally proposed in [13], [27], followed by numerous modifications reflecting, e.g., model

Manuscript received December 12, 2016; revised December 12, 2016. First published December 12, 2016; current version published December 12, 2016. The work of K. Dedecius was supported by the Czech Science Foundation, grant No. 14–06678P. The work of P. M. Djurić was supported by NSF under Awards CCF-1320626 and CCF-1618999. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Kostas Berberidis.

K. Dedecius is with the Institute of Information Theory and Automation, Czech Academy of Sciences, Pod Vodárenskou věží 4, 182 08 Prague 8, Czech Republic (email: dedecius@utia.cas.cz).

P.M. Djurić is with the Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY 11794-2350 USA (e-mail: petar.djuric@stonybrook.edu).

Digital Object Identifier XXX/XXX

sparsity [28] and node-specific parameters [29], [30]. Another group of solutions relies on the recursive least-squares (RLS) method, with the basic diffusion RLS being reported in [12] and its modified version with partial diffusion saving communication resources in [17]. In the field of linear state-space models, the diffusion Kalman filters were proposed in [31]. During the combination phase, these filters fuse the local point estimates by leaving the connected covariance matrices intact. Indeed, this saves communication resources, but may lead to inconsistent estimates. This issue was addressed by a covariance intersection-based combination phase [32]. The way towards stochastic optimization with non-smooth regularizers applicable, e.g., to logistic regression, was recently introduced too. In [33], the authors consider data with an unknown distribution, and proceed with non-smooth regularization of the expected loss function in the role of a risk function. Naturally, there are many other diffusion algorithms, and they consider inference of mixtures [34]–[36] or use particle filters [24], [37], [38].

A common feature of the existing methods is their independent derivation from a particular traditional estimation method, e.g., RLS and LMS estimation or Kalman filtering. Recognizing the common principles of inference, we propose to adopt the probabilistic Bayesian paradigm, providing theoretically consistent yet highly versatile methods. The formulation of the diffusion framework in this scope, originally proposed by the first author in [16], then yields a generic Bayesian diffusion estimator, such that (i) its application to a wide class of models is straightforward and under certain conditions it is analytically obtained, and (ii) its use in conjunction with many popular Bayesian techniques is straightforward without modifications including forgetting techniques for tracking slowly varying parameters [16] and dynamic model averaging [39].

We stress that this paper aims at formulating a general Bayesian framework for diffusion estimation, naturally yielding basic diffusion-based estimation methods as special cases. We work with models based on the exponential family of distributions. The paper does not aspire to develop methods competing with particular problem-oriented diffusion algorithms like the colored-noise RLS [14], node-specific parameter estimation [30] or the like. The paper significantly extends earlier results of one of the authors, devoted to specific generalized linear models cases [16], [40], [41]. In particular, it proposes a consolidated Bayesian framework for the diffusion inference of a wide class of models, and provides a consistent explanation of its derivation and properties.

II. Preliminaries on Bayesian estimation

Consider discrete-time sequential modeling of a stochastic process with observations $y_t, t = 1, 2, \ldots$ determined by an unknown parameter θ and, if present, an explanatory variable z_t, e.g., the regressor. First, we assume that θ is fixed and later in Sections VIII and IX where we address hidden Markov models and slowly time-varying parameters, respectively, it is relaxed.

The statistical approach to modeling portrays the true observations-generating system by a model represented by a probability distribution with a density $f(y_t|z_t, \theta)$. If z_t is not present, we have $f(y_t|z_t, \theta) \equiv f(y_t|\theta)$. Whatever the goal of modeling is, be it forecasting, smoothing or filtering, a reliable determination of the value of θ is of paramount importance.

The present paper adheres to the sequential Bayesian framework, where one estimates the unknown parameter θ by exploiting a prior distribution $\pi(\theta|y_{0:t-1}, z_{0:t-1})$, which quantifies the accumulated knowledge about θ from the past observations $y_{0:t-1} = \{y_0, \ldots, y_{t-1}\}$ and regressors $z_{0:t-1} = \{z_0, \ldots, z_{t-1}\}$. The values y_0 and z_0 can be viewed as pseudo-observations, expressing any available knowledge at the beginning of modeling, including total ignorance. The Bayes' theorem then sequentially incorporates the newly observed y_t and z_t via

$$\pi(\theta|y_{0:t}, z_{0:t}) \propto f(y_t|z_t, \theta)\pi(\theta|y_{0:t-1}, z_{0:t-1}). \qquad (1)$$

In writing the above equation, we assume $f(y_t|y_{0:t-1}, z_{0:t}, \theta) = f(y_t|z_t, \theta)$, which implies that the observation y_t is independent of previous observations given the latest explanatory variables z_t and the parameters θ. This is a standard assumption in statistical signal modeling. As already outlined, in the last part of the paper, we present four examples, and for each of them, we show the validity of this assumption. We also point out that if this assumption cannot be used, the methodology presented in the sequel still holds but with appropriate modification in notation.

A thorough inspection of (1) reveals a feature crucial for the following development: the observations can be incorporated as a batch, that is, for any positive $\tau \leq t$,

$$\pi(\theta|y_{0:t}, z_{0:t}) \propto \pi(\theta|y_{0:\tau-1}, z_{0:\tau-1}) \prod_{\tilde{\tau}=\tau}^{t} f(y_{\tilde{\tau}}|z_{\tilde{\tau}}, \theta). \qquad (2)$$

The resulting knowledge about θ is described by the posterior distribution, whose statistics may serve as point estimates. Mostly, the mean is preferred, but the mode or median are frequently used too [42].

The greatest hindrance of the Bayesian approach lies in its rare cases of analytical and/or computationally low-cost numerical tractability of posterior distributions. Still, if the model $f(y_t|z_t, \theta)$ belongs to the exponential family of distributions [43], there is a way towards analytical results.

Definition 1 (Exponential family of distributions)**.** *Assume a random variable y conditioned by the variable z and the parameter θ. The exponential family of distributions is a class of distributions with probability density functions of the form*

$$f(y|z, \theta) = h(y, z)g(\theta) \exp\left[\eta^{\mathsf{T}} T(y, z)\right],$$

where $\eta \equiv \eta(\theta)$ is the natural parameter, $T(y, z)$ is a sufficient statistic of a fixed size, $h(y, z)$ is a known function, and $g(\theta)$ is a known normalizing function. If $\eta(\theta) = \theta$, the family is called canonical.

The definition requires vectorization of all matrices included in the argument of the exponential function. However, to prevent confusion, a popular approach is to rearrange the relevant terms using the *trace* operator preserving these matrices, and to call the resulting *unvectorized* terms the natural parameters

η and sufficient statistics $T(\cdot)$ too. The result is a more comprehensible "unvectorized" version of the standard Definition 1. The normal distribution (13) is a classical example.

Several important distributions belong to the exponential family including the normal, beta, gamma, multinomial, and Poisson distributions. The exponential family distributions are closely related to the conjugate prior distributions of θ, making the Bayesian updates (1) and (2) analytical [44].

Definition 2. *Assume that y given z and θ follows an exponential family distribution. The conjugate prior distribution of θ is a distribution with a probability density function of the form*

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

where ξ is a hyperparameter of the same size as $T(y, z)$, $\nu \in \mathbb{R}^{+}$ is a scalar hyperparameter, and $q(\xi, \nu)$ is a known function. The symbol $g(\theta)$ is the same function as in the exponential family distribution.

It is straightforward to see that under conjugacy, the Bayesian update (1) takes the form of a simple prior hyperparameters update, i.e.,

$$\xi_t = \xi_{t-1} + T(y_t, z_t), \quad \text{and} \quad \nu_t = \nu_{t-1} + 1. \tag{3}$$

The update corresponding to (2) has the form

$$\xi_t = \xi_{\tau-1} + \sum_{\widetilde{\tau}=\tau}^{t} T(y_{\widetilde{\tau}}, z_{\widetilde{\tau}}), \quad \text{and} \quad \nu_t = \nu_{\tau-1} + t - \tau + 1.$$

It will be shown that the functional form of the conjugate prior also provides analytically tractable merging of posterior distributions.

III. Estimation by diffusion

We consider a network represented by a connected undirected graph consisting of a set of nodes (vertices) $\mathcal{I} = \{1, \ldots, I\}$. The nodes are linked by a set of edges which determine the topology of the network. Each node $i \in \mathcal{I}$ directly communicates only with the nodes from its *neighborhood* (nodes that share edges with i). We denote this set by $\mathcal{I}_i$ and assume that it includes the node i too. The estimation by diffusion runs by exchanging measurements between the nodes during an *adaptation* phase and exchanging estimates during a *combination* phase. These phases will be described below in terms of Bayesian probability theory. We discriminate four different schemes: one that employs adaptation only (A), one with combination only (C), and two that use the two phases but in different orders. We refer to the latter two as adapt-then-combine (ATC) and combine-then-adapt (CTA) schemes. Table I summarizes the available diffusion strategies.

A. Adaptation phase

Assume that the nodes $i \in \mathcal{I}$ observed $y_{i,t}$ and know $z_{i,t}$. The aim of the adaptation phase is to enrich the statistical knowledge of each node by incorporation of the neighbors' observations. In a Bayesian context, this means to perform a batch update similarly to (2) with a fixed time index as in a static (non-sequential) case. Fixing a node $i \in \mathcal{I}_i$ and denoting its prior density $\pi_i(\theta|\zeta_{i,t-1})$, where $\zeta_{i,t-1}$ represents all the information available to node i by time $t-1$, which includes its own past observations and those of its neighbors, as well as the parameters of all previous posteriors of its neighbors, the update takes the form

$$\begin{aligned} \pi_i(\theta|\zeta_{i,t}) &= \pi_i(\theta|\zeta_{i,t-1}, \bar{y}_{i,t}, \bar{z}_{i,t}) && (4) \\ &\propto \pi_i(\theta|\zeta_{i,t-1}) \prod_{j \in \mathcal{I}_i} \left[f(y_{j,t}|z_{j,t}, \theta)\right]^{c_{ij,t}}, && (5) \end{aligned}$$

where the "bar" notation in (4) refers to all the new observations and explanatory variables available to node i at time t, $c_{ij,t} \in \{0, 1\}$ are adaptation weights assigned by the node i to its neighbors $j \in \mathcal{I}_i$. If the observation $y_{j,t}$ is not an outlier, then $c_{ij,t} = 1$; otherwise, $c_{ij,t} = 0$. The hyperparameters of the conjugate prior densities are updated as follows:

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

Obviously, the communication requirements of the adaptation phase are directly determined by the size of the sufficient statistic $T(y_{j,t}, z_{j,t})$. Provided that before vectorization – see comment below Definition 1 – this statistic is described by an $M \times N$ matrix of floating point numbers, then a node i needs to obtain $(|\mathcal{I}_i| - 1)MN$ floating point numbers, where $|\cdot|$ denotes cardinality of the argument.

B. Combination phase

The purpose of the combination phase is to share and collaboratively improve — *optimally combine* — the estimates of the nodes. If the combination phase follows after the adaptation phase, i.e., the *adapt-then-combine* (ATC) strategy is adopted, the estimates are represented by the local posterior distributions $\pi_i(\theta|\zeta_{i,t})$ resulting from (5). Alternatively, the *combine-then-adapt* (CTA) strategy combines the local prior distributions $\pi_i(\theta|\zeta_{i,t-1})$ before the adaptation phase. Below, we describe the combination phase for both strategies, though the rest of the paper puts emphasis on the ATC strategy.

In Bayesian theory, a measure for optimality that is often advocated is the Kullback-Leibler divergence (KLD). The goal is to find such a final density $\widetilde{\pi}_i(\theta|\zeta_{i,t})$ (or $\widetilde{\pi}_i(\theta|\zeta_{i,t-1})$) whose divergence from the densities $\pi_j(\theta|\zeta_{j,t})$ (or $\pi_j(\theta|\zeta_{j,t-1})$) of all the neighbors $j \in \mathcal{I}_i$ is minimal. The following proposition gives a solution of this task.

Proposition 1. *Given a node $i \in \mathcal{I}$, densities $\pi_j(\theta|\cdot)$ of its neighbors $j \in \mathcal{I}_i$ and unit $|\mathcal{I}_i|$-simplex weights $a_{ij,t}$ expressing the degree of belief of i in the j's information, the density $\widetilde{\pi}_i(\theta|\cdot)$ combining $\pi_j(\theta|\cdot)$ in the Kullback-Leibler optimal sense and minimizing the cumulative loss*

$$\sum_{j \in \mathcal{I}_i} a_{ij,t} \mathcal{D}\left(\widetilde{\pi}_i\left(\theta|\cdot\right) \middle|\middle| \pi_j\left(\theta|\cdot\right)\right), \tag{7}$$

where the KLD is given by

$$\mathcal{D}\left(\widetilde{\pi}_i\left(\theta|\cdot\right) \middle|\middle| \pi_j\left(\theta|\cdot\right)\right) = \mathbb{E}_{\widetilde{\pi}_i}\left[\log \frac{\widetilde{\pi}_i\left(\theta|\cdot\right)}{\pi_j\left(\theta|\cdot\right)}\right],$$

has the form

$$\widetilde{\pi}_i(\theta|\cdot) \propto \prod_{j\in\mathcal{I}_i} [\pi_j(\theta|\cdot)]^{a_{ij,t}} .$$

Proof. Using the definition of the KLD and properties of expectation and logarithm it follows that

$$\sum_{j\in\mathcal{I}_i} a_{ij,t}\mathcal{D}\left(\widetilde{\pi}_i(\theta|\cdot)\,||\pi_j(\theta|\cdot)\right) = \sum_{j\in\mathcal{I}_i} a_{ij,t}\mathbb{E}_{\widetilde{\pi}_i}\left[\log\frac{\widetilde{\pi}_i(\theta|\cdot)}{\pi_j(\theta|\cdot)}\right]$$

$$=\mathbb{E}_{\widetilde{\pi}_i}\left[\log\frac{\widetilde{\pi}_i(\theta|\cdot)}{c\prod_{j\in\mathcal{I}_i}[\pi_j(\theta|\cdot)]^{a_{ij,t}}}\right]=\mathcal{D}\left(\widetilde{\pi}_i(\theta|\cdot)\,\middle|\middle|\,c\prod_{j\in\mathcal{I}_i}[\pi_j(\theta|\cdot)]^{a_{ij,t}}\right),$$

where c is a proportionality constant. From the definition of the KLD, convexity of $-\log(\cdot)$, and the Jensen's inequality, it follows that

$$\begin{aligned}\mathcal{D}\left(\widetilde{\pi}_i(\theta|\cdot)\,\middle|\middle|\,c\prod_{j\in\mathcal{I}_i}[\pi_j(\theta|\cdot)]^{a_{ij,t}}\right) &= \mathbb{E}_{\widetilde{\pi}_i}\left[\log\frac{\widetilde{\pi}_i(\theta|\cdot)}{c\prod_{j\in\mathcal{I}_i}[\pi_j(\theta|\cdot)]^{a_{ij,t}}}\right]\\ &= \mathbb{E}_{\widetilde{\pi}_i}\left[-\log\frac{c\prod_{j\in\mathcal{I}_i}[\pi_j(\theta|\cdot)]^{a_{ij,t}}}{\widetilde{\pi}_i(\theta|\cdot)}\right]\\ &\geq -\log\mathbb{E}_{\widetilde{\pi}_i}\left[\frac{c\prod_{j\in\mathcal{I}_i}[\pi_j(\theta|\cdot)]^{a_{ij,t}}}{\widetilde{\pi}_i(\theta|\cdot)}\right] = -\log 1 = 0.\end{aligned}$$

That is, the KLD is nonnegative and minimal (zero), if

$$\widetilde{\pi}_i(\theta|\cdot) \propto \prod_{j\in\mathcal{I}_i} [\pi_j(\theta|\cdot)]^{a_{ij,t}} \qquad \textit{a.e.},$$

which concludes the proof. □

Note that here we only proved the form of the optimal posterior. In other words, we know what it is if we know the coefficients a_{ij}. How we find the best set of coefficients is addressed in Section V.

An immediate consequence of Proposition 1 is the simplicity of the combination of conjugate priors. Namely, if the ATC strategy is adopted, we can write,

$$\widetilde{\xi}_{i,t} = \sum_{j\in\mathcal{I}_i} a_{ij,t}\xi_{j,t}, \quad \text{and} \quad \widetilde{\nu}_{i,t} = \sum_{j\in\mathcal{I}_i} a_{ij,t}\nu_{j,t}. \tag{8}$$

Alternatively, if the CTA strategy is used, the combination of conjugate priors takes the form

$$\widetilde{\xi}_{i,t-1} = \sum_{j\in\mathcal{I}_i} a_{ij,t}\xi_{j,t-1}, \quad \text{and} \quad \widetilde{\nu}_{i,t-1} = \sum_{j\in\mathcal{I}_i} a_{ij,t}\nu_{j,t-1}. \tag{9}$$

In the case of the ATC strategy, the optimal density $\widetilde{\pi}_i(\theta|\zeta_{i,t})$ is the final estimation product at the time instant t. Subsequently, at the next time step, it serves as the prior distribution $\pi_i(\theta|\zeta_{i,t-1})$ in (5). On the other hand, in the CTA strategy, $\widetilde{\pi}_i(\theta|\zeta_{i,t-1})$ does not contain the t-time observations yet. It enters the adaptation step (5) where it is identified with $\pi_i(\theta|\zeta_{i,t-1})$.

The communication load of the combination phase at the ith node is equal to that of the adaptation phase plus the need to transmit $|\mathcal{I}_i| - 1$ floating point numbers required for $\nu_{j,t}, j \in \mathcal{I}_i$. There are cases where the νs are not needed for inference, as is the case for the Kalman filter, Section VIII.

Strategy	First step	Second step
ATC	Adaptation (6)	Combination (8)
CTA	Combination (9)	Adaptation (6)
A	Adaptation (6)	—
C	Adaptation (6) with $c_{ij,t} = \mathbb{1}_{[i=j]}$	Combination (8)

TABLE I

DIFFUSION STRATEGIES: ADAPT-THEN-COMBINE (ATC), COMBINE-THEN-ADAPT (CTA), ADAPTATION-ONLY (A), AND COMBINATION-ONLY (C) AS A SPECIAL CASE OF ATC.

Finally, we remark that the Kullback-Leibler optimality criterions are widespread in the information fusion literature. For instance, in the consensus probability hypothesis density filters [45] and [46] it is approached during an *iterative* consensus step.

IV. Properties of diffusion estimators

This section discusses some important properties of the diffusion estimator. The time indexing corresponds with the ATC strategy.

A. Asymptotics of the adaptation phase

The asymptotic properties of the adaptation phase are fully in agreement with the consistency of Bayes' estimators, underpinned by the law of large numbers ("more observations lead to better estimates"). Generally, if $g(y|z)$ is a true observations-generating density and $f(y|z,\theta)$ is a model chosen to approximate it, then the estimation goal is to find $\hat{\theta}$ such that

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \mathcal{D}\left(g(y|z)||f(y|z,\theta)\right), \tag{10}$$

where Θ is a compact set. Simply put, the consistency of the Bayesian estimator states that with the growing number of observations, the posterior point estimates converge to the value $\hat{\theta}$ that minimizes (10), or to a pseudo-true parameter that minimizes the divergence under model imprecisions [51]. The purpose of the adaptation phase is to accelerate this convergence by increasing the number of observations. Note also the conceptual similarity of (10) and (7).

Since the adaptation step is in accordance with the classical asymptotic properties of Bayesian inference, the reader is referred to relevant literature [47], [49], [51], [52].

B. Asymptotics of the combination phase

Recall that during the combination phase the nodes combine posterior densities provided by their neighbors. This entails that this information diffuses through the network, unlike the information in the measurements exchanged during the adaptation phase.

In order to examine the combination phase properties, we adopt the following three assumptions:

1) The operational conditions are regular in the sense that there is no error associated with the nodes or communication links, e.g., due to their malfunctions. This assumption can be removed by appropriate tuning of $a_{ij,t}$ and $c_{ij,t}$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2641380, IEEE Transactions on Signal Processing

2) The estimation conditions are regular in the sense that the observations follow the considered models and generally allow estimation of their parameters.
3) The initial prior distributions $\pi_i(\theta|\bar{y}_0, \bar{z}_0)$ are (weakly) informative, i.e., there is no misleading or convergence-preventing prior information. *Ipso facto*, the large sample theory guarantees estimates convergence even under misleading prior distribution if its support covers the true parameter value. This technical assumption, routinely applied in practice, assures that the posterior distribution *correctly* quantifies the state of knowledge about θ.

First, we recall a proposition on asymptotic posterior normality from [48] (see also [49]). We assume the following conditions to hold:

C1 As $t \to \infty$, the largest eigenvalue of Σ_t tends to zero.

C2 For any $\epsilon > 0$, there exists T and $\delta > 0$ such that for any $t > T$ and $\theta \in B_\delta(\widehat{\theta}_t) = \{\theta \in \Theta; |\theta - \widehat{\theta}_t| < \delta\}$ (with Θ being the support of θ), $\Sigma_t^{-1}(\theta) = \frac{\partial^2 \log \pi(\theta|y_{0:t}, z_{0:t})}{\partial\theta\partial\theta^\mathsf{T}}$ exists and satisfies

$$\mathsf{I} - A(\epsilon) \le \Sigma_t^{-1}(\theta)\Sigma_t(\widehat{\theta}_t) \le \mathsf{I} + A(\epsilon)$$

where I is a $d_\theta \times d_\theta$ identity matrix and $A(\epsilon)$ is a $d_\theta \times d_\theta$ symmetric positive semidefinite matrix whose largest eigenvalue tends to zero as $t \to \infty$.

C3 For any $\delta > 0$, $\int_{B_\delta(\widehat{\theta}_t)} \pi(\theta|y_{0:t}, z_{0:t})\mathrm{d}\theta \to 1$ as $t \to \infty$.

Proposition 2. *For each t, consider $\pi(\theta|y_{0:t}, z_{0:t})$ as the density function of a random quantity θ and define, $\phi_t = \Sigma_t^{-1}(\theta - \widehat{\theta}_t)$. Then given* C1 *and* C2, C3 *is a necessary and sufficient condition for the distribution of ϕ to converge to $f(\phi) = (2\pi)^{-d_\theta/2} e^{-\frac{1}{2}\phi^\mathsf{T}\phi}$.*

Proof: See [48].

Asymptotic normality of the posterior distribution of the constant parameter θ is assumed in the sequel.

Next, we present a proposition about the KLD between the posteriors of a fusion center and a noncooperative node.

Proposition 3. *Let $\pi_C(\theta|\zeta_t) = \pi_C(\theta|y_{1,0:t}, \dots, y_{I,0:t}, z_{1,0:t}, \dots, z_{I,0:t})$ be the normal posterior density of a fusion center that receives the sufficient statistics of all the nodes in a network of two or more nodes, and $\pi_i(\theta|\zeta_{i,t}) = \pi_i(\theta|y_{i,0:t}, z_{i,0:t})$ the normal posterior density of a noncooperative node i. Assume that the Fisher information matrix of each node in the network is the same. Then, the Kulback-Leibler distance between $\pi_C(\theta|\zeta_t)$ and $\pi_i(\theta|\zeta_{i,t})$ remains finite as $t \to \infty$.*

Proof: The Kulback-Leibler distance between $\pi_C(\theta|\zeta_t)$ and $\pi_i(\theta|\zeta_{i,t})$ is given by

$$\mathcal{D}(\pi_C(\theta|\zeta_t)||\pi_i(\theta|\zeta_{i,t}) = \frac{1}{2}\left(\mathrm{tr}\left(\Sigma_{c,t}^{-1}\Sigma_{i,t}\right) + \ln\left(\frac{|\Sigma_{c,t}|}{|\Sigma_{i,t}|}\right)\right.$$
$$\left. -d_\theta + \left(\widehat{\theta}_{c,t} - \widehat{\theta}_{i,t}\right)^\mathsf{T} \Sigma_{c,t}^{-1}\left(\widehat{\theta}_{c,t} - \widehat{\theta}_{i,t}\right)\right).$$

When $t \to \infty$, we have [48]

$$\Sigma_{c,t} \quad \to \quad ItF(\widehat{\theta}_{c,t}),$$

where $F(\widehat{\theta}_{c,t})$ is the Fisher information matrix, and I is the number of nodes in the network. Similarly,

$$\Sigma_{i,t} \quad \to \quad tF(\widehat{\theta}_{i,t}),$$

Under the conditions C1-C3, we have that both $\widehat{\theta}_{c,t}$ and $\widehat{\theta}_{i,t}$ tend to θ as $t \to \infty$. Let

$$\widehat{\theta}_{i,t} - \theta \quad < \quad \frac{1}{t^\epsilon}\mathbb{1},$$

for some $\epsilon > 0$ and where $\mathbb{1} = [1\ 1\ \cdots\ 1]^\mathsf{T}$. Then,

$$\widehat{\theta}_{c,t} - \theta \quad < \quad \frac{1}{(It)^\epsilon}\mathbb{1}.$$

Finally, if we use $F(\widehat{\theta}_{i,t}) \approx F(\widehat{\theta}_{c,t})$ for very large t, we can readily show that

$$\lim_{t\to\infty} \mathcal{D}(\pi_C(\theta|\zeta_t)||\pi_i(\theta|\zeta_{i,t}) = \frac{1}{2}\left(\frac{d_\theta}{I} + d_\theta \ln I - d_\theta\right).$$

When $I > 1$, this distance is greater than zero. □

Now we state a proposition which maintains that under certain conditions the KLD between the posteriors of a fusion center and a cooperative node, respectively, tends to zero. We prove the proposition for a network where the nodes implement only the combination scheme.

Proposition 4. *Let the conditions* C1-C3 *hold. Assume further that the coefficients $a_{ij,t}$ are* not *functions of time, that the network is fully connected, and that the weight matrix A with elements a_{ij} satisfies the following three conditions:*

C4 $\mathbb{1}^\mathsf{T}\mathsf{A} = \mathbb{1}^\mathsf{T}$,
C5 $\mathsf{A}\mathbb{1} = \mathbb{1}$,
C6 $\rho\left(\mathsf{A} - (1/I)\mathbb{1}\mathbb{1}^\mathsf{T}\right) < 1$,

where $\rho(\cdot)$ is the spectral radius of the argument. Then

$$\lim_{t\to\infty} \mathcal{D}\left(\pi_C(\theta|\zeta_t)||\pi_i(\theta|\zeta_{i,t})\right) \quad \to \quad 0.$$

We prove the proposition by following the proof of a related theorem from [50]. Before we proceed, we define the precision matrices $\Lambda_{i,t}$ and $\Lambda_{c,t}$ by $\Lambda_{i,t} = \Sigma_{i,t}^{-1}$ and $\Lambda_{c,t} = \Sigma_{c,t}^{-1}$. First, we need three lemmas.

Lemma 1. *Let A in Proposition 3 be defined by conditions* C4-C6 *and let the conditions* C1-C3 *hold. Let also*

$$\begin{aligned} \Delta_{i,t} &= \Lambda_{i,t} - \Lambda_{c,t}, \\ \delta_{i,t} &= \widehat{\theta}_{i,t} - \widehat{\theta}_{c,t}. \end{aligned}$$

Then for all $i \in \mathcal{I}$ and for all $t \ge 0$, the elements of $\Delta_{i,t}$ and $\delta_{i,t}$ are bounded.

Proof: The proof is analogous to the proof of Lemma 1 in [50], and so we do not repeat it here.

Lemma 2. *Let the conditions* C1-C6 *hold. Then*

$$\begin{aligned} \lim_{t\to\infty} \Lambda_{c,t}^{-1} &= 0_{d_\theta \times d_\theta}. \\ \lim_{t\to\infty} \Lambda_{i,t}^{-1} &= 0_{d_\theta \times d_\theta}, \end{aligned}$$

where $0_{d_\theta \times d_\theta}$ stands for a $d_\theta \times d_\theta$ matrix with elements equal to zero.

Proof: The proof follows the same lines as those from the proof of Lemma 2 in [50], and we omit it here.

The next lemma is a straightforward application of Proposition 1 to normal densities.

Lemma 3 (Combination of normal densities)**.** *Let $\pi_j(\theta|\cdot)$ be n-variate normal densities with column mean vectors $\mu_{j,t} \in \mathbb{R}^n$ and covariance matrices $\Sigma_{j,t} \in \mathbb{R}^{n\times n}$, $j \in \mathcal{I}_i$. Let $a_{ij,t}$ be positive weights taking values in the unit $|\mathcal{I}_i|$-simplex. Then, the density*

$$\widetilde{\pi}_i(\theta|\cdot) \propto \prod_{j\in\mathcal{I}_i} [\pi_j(\theta|\cdot)]^{a_{ij,t}} \tag{11}$$

combined according to Proposition 1 is again a normal density with a mean vector and covariance matrix given by

$$\widetilde{\mu}_{i,t} = \widetilde{\Sigma}_{i,t}\left(\sum_{j\in\mathcal{I}_i} a_{ij,t}\Sigma_{j,t}^{-1}\mu_{j,t}\right) \text{ and } \widetilde{\Sigma}_{i,t} = \left[\sum_{j\in\mathcal{I}_i} a_{ij,t}\Sigma_{j,t}^{-1}\right]^{-1}. \tag{12}$$

Proof. Let us drop the time indices. The normal density $\pi_j(\theta|\cdot)$ can be written in the exponential family form

$$\begin{aligned}\pi_j(\theta|\mu_j,\Sigma_j) &= (2\pi)^{\frac{-n}{2}}(\det\Sigma_j)^{\frac{-1}{2}} e^{\frac{-1}{2}(\theta-\mu_j)^\intercal\Sigma_j^{-1}(\theta-\mu_j)} \\ &\propto \exp\left\{\operatorname{Tr}\left(\begin{bmatrix}\mu_j^\intercal\Sigma_j^{-1} \\ -\frac{1}{2}\Sigma_j^{-1}\end{bmatrix}^\intercal \begin{bmatrix}\theta^\intercal \\ \theta\theta^\intercal\end{bmatrix}\right) - \frac{1}{2}\mu_j^\intercal\Sigma_j^{-1}\mu_j\right\} \end{aligned}\tag{13}$$

with

$$\eta_j = \begin{bmatrix}\mu_j^\intercal\Sigma_j^{-1} \\ -\frac{1}{2}\Sigma_j^{-1}\end{bmatrix} \quad\text{and}\quad T(\theta) = \begin{bmatrix}\theta^\intercal \\ \theta\theta^\intercal\end{bmatrix}.$$

The weighted geometric product (11) with θ fixed leads to

$$\widetilde{\eta}_i = \begin{bmatrix}\widetilde{\mu}_i^\intercal\widetilde{\Sigma}_i^{-1} \\ -\frac{1}{2}\widetilde{\Sigma}_i^{-1}\end{bmatrix} = \sum_{j\in\mathcal{I}_i} a_{ij}\eta_j = \sum_{j\in\mathcal{I}_i} a_{ij}\begin{bmatrix}\mu_j^\intercal\Sigma_j^{-1} \\ -\frac{1}{2}\Sigma_j^{-1}\end{bmatrix},$$

from which the claimed result follows. □

This result is also known as the covariance intersection.

Proof of Proposition 3: The Kullback-Leibler distance will satisfy (11) if

$$\lim_{t\to\infty}\left(\widehat{\theta}_{c,t} - \widetilde{\theta}_{i,t}\right) = 0_{d_\theta\times 1}, \tag{14}$$

$$\lim_{t\to\infty}\Lambda_{c,t}\widetilde{\Lambda}_{i,t}^{-1} = 0_{d_\theta\times d_\theta}, \tag{15}$$

where $\widetilde{\theta}_{i,t}$ is the estimate of the ith node after the combination step and $0_{d_\theta\times 1}$ is a $d_\theta \times 1$ vector of zeros.

We now rewrite (12) as

$$\widetilde{\theta}_{i,t} = \widetilde{\Lambda}_{i,t}^{-1}\left(\sum_{j\in\mathcal{I}_i} a_{ij}\Lambda_{j,t}\widehat{\theta}_{j,t}\right),$$

where $\widehat{\theta}_{i,t}$ is the estimate of the ith node at t after adaptation by using only its own observations. Next we use Lemma 1 and write

$$\begin{aligned}\widetilde{\theta}_{i,t} &= (\Lambda_{c,t} + \widetilde{\Delta}_{i,t})^{-1} \\ &\times \sum_{j\in\mathcal{I}_i} a_{ij}(\Lambda_{c,t} + \Delta_{j,t})(\widehat{\theta}_{c,t} + \delta_{j,t})\end{aligned}$$

According to the matrix inversion lemma,

$$(\Lambda_{c,t} + \widetilde{\Delta}_{i,j})^{-1} = \Lambda_{c,t}^{-1} - \Lambda_{c,t}^{-1}(\Lambda_{c,t}^{-1} + \widetilde{\Delta}_{i,t}^{-1})^{-1}\Lambda_{c,t}^{-1},$$

and by using Lemmas 1 and 2, we immediately have

$$\lim_{t\to\infty}\Lambda_{c,t}^{-1}(\Lambda_{c,t}^{-1} + \widetilde{\Delta}_{i,t}^{-1})^{-1} = 0_{d_\theta\times d_\theta},$$

and that all the elements of $\Lambda_{c,t}^{-1}\delta_{i,t}$ converge to a zero vector. This proves the condition (14).

We prove the condition (15) by using

$$\Lambda_{c,t}\Lambda_{i,t}^{-1} = \Lambda_{c,t}^{-1}(\Lambda_{c,t} + \Delta_{i,t})^{-1},$$

and

$$\lim_{t\to\infty}\Lambda_{c,t}^{-1}\Delta_{i,t} = 0_{d_\theta\times d_\theta},$$

which follows from Lemmas 1 and 2. The last two equations prove (15). □

C. ATC versus CTA estimators

The literature on diffusion estimation often proposes both the adapt-then-combine (ATC) and the combine-then-adapt (CTA) strategies, though it has been repeatedly shown that the ATC-based algorithms outperform the CTA-based ones [3]. The Bayesian paradigm allows to prove this fact abstractly, independently of any particular model case. It suffices to show that the ATC estimator involves more observations than its CTA counterpart; the rest then follows from the outlined asymptotic properties.

Let us omit time indices and the explanatory variables for simplicity. Without a loss of generality, assume that the network nodes $i \in \mathcal{I}$ start with the same prior distribution $\pi_i(\theta)$, $c_{ij} = 1$ for all i and j, and that the weights a_{ij} are assigned. The ATC estimator at i combining the adapted posterior densities yields the following density:

$$\begin{aligned}\widetilde{\pi}_i(\theta|\bar{y}) &\propto \prod_{j\in\mathcal{I}_i}\left[\pi_j\left(\theta|\{y_k\}_{k\in\mathcal{I}_j}\right)\right]^{a_{ij}} \\ &\propto \prod_{j\in\mathcal{I}_i}\left[\pi_j(\theta)\prod_{k\in\mathcal{I}_j}[f(y_k|\theta)]^{c_{ij}}\right]^{a_{ij}} \\ &\propto \pi_i(\theta)\prod_{\substack{j\in\mathcal{I}_i \\ k\in\mathcal{I}_j}}[f(y_k|\theta)]^{a_{ij}}. \end{aligned}\tag{16}$$

The CTA estimator produces the following density:

$$\begin{aligned}\widetilde{\pi}_i(\theta|\bar{y}) &\propto \prod_{j\in\mathcal{I}_i}[\pi_j(\theta)]^{a_{ij}}\prod_{j\in\mathcal{I}_i}[f(y_j|\theta)]^{c_{ij}} \\ &\propto \pi_i(\theta)\prod_{j\in\mathcal{I}_i} f(y_j|\theta). \end{aligned}\tag{17}$$

The comparison of (16) and (17) reveals that both the ATC and CTA estimators involve observations from the neighbors of i. However, the former additionally involves observations from all the neighbors of the neighbors.

D. Combination-only and adaptation-only estimation

An interesting case occurs if $c_{ij,t} = 1_{[i=j]}$ in ATC, leading to

$$\widetilde{\pi}_i(\theta|\zeta_t) \propto \prod_{j\in\mathcal{I}_i} \left[\pi_j(\theta|\zeta_{t-1}) f(y_{j,t}|z_{j,t},\theta)\right]^{a_{ij,t}},$$

from which it follows that at time t, the estimator at node i updates the posterior by using its observations only. That is, the adaptation phase is replaced by a local Bayesian update. Compared to the ATC, the observations of the other nodes in the network are available to i through the neighbors' posterior densities with a time delay equal to the number of hops to these nodes minus 1. This means that the ATC and combination-only estimators are asymptotically equivalent. This can be useful in practice, e.g., in processing big data sets to avoid the adaptation phase, which saves on communication resources. This phenomenon will be demonstrated in the examples.

The adaptation-only estimation sets $a_{ij,t} = \mathbb{1}_{[i=j]}$, that is, the combination step is skipped. The inference at node i reduces to the ordinary Bayesian estimation with $|\mathcal{I}_i|$ observations from the neighborhood. Indeed, the information about the estimates of θ does not diffuse throughout the network.

For completeness we remark that setting $a_{ij,t} = c_{ij,t} = \mathbb{1}_{[i=j]}$ leads to the non-cooperative ordinary Bayesian inference from the local observations only.

E. Numerical comparison of strategies

In order to compare the diffusion strategies numerically, we assume estimation of a common normal mean by a simple network of 6 nodes depicted in Figure 1 (left). The nodes observed samples from $\mathcal{N}(10, \sigma_i^2)$ whose standard deviations σ_i are depicted in Figure 1 (right). Six possible scenarios were compared: centralized, where all the data were processed by a single node, adapt-then-combine (ATC), combine-then-adapt (CTA), adaptation-only (A), combination-only (C) and a strategy where the nodes did not cooperate at all (No coop.). The simulations started from flat normal prior distributions $\mathcal{N}(0, 1000)$ and the posterior means and variances were averaged over the network. The combination weights were static and uniform, $a_{ij} = |\mathcal{I}_i|^{-1}$. The resulting posterior estimates of means $\widehat{\mu}$, and variances of these estimates represented by $\widehat{\mu} \pm 3$ standard deviations are depicted in Figure 2 for time instants $t = 1$ (top) and $t = 3$ (bottom). The results were averaged over 100 experiment runs.

The results are in accordance with the analyses conducted in the previous sections. The difference between ATC and CTA vanishes with the increasing number of incorporated observations and their performances become close to the performance of the centralized strategy. The combination step assures gradual diffusion of information in the network, which is particularly demonstrated by the combination-only scenario (C). In the adaptation-only (A) strategy, the information (in the observations) does not diffuse further than 1 hop. Hence, although the convergence is faster than in the non-cooperation strategy, it is slower than in other strategies.

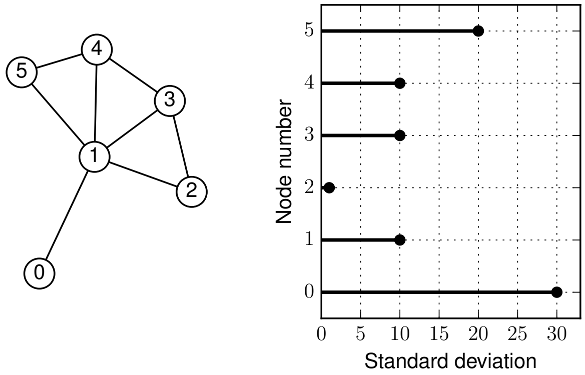


Fig. 1. Network topology (left) and standard deviations of nodes observations (right).

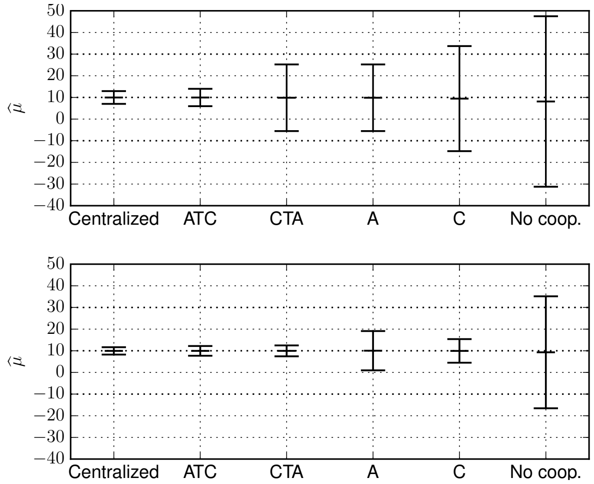


Fig. 2. Posterior point estimates of means averaged over the network after 1 (top figure) and 3 (bottom figure) time steps. The intervals around $\widehat{\mu}$ show averaged variances in terms of ± 3 standard deviations.

V. Determination of combiner weights

There are two different sets of weights in the proposed algorithm: the adaptation weights $c_{ij,t}$ and the combination weights $a_{ij,t}$, both taking values in the interval $[0, 1]$, but the latter obeying an additional requirement of summing to unity.

A. Adaptation weights $c_{ij,t}$

The meaning of $c_{ij,t}$ in this paper is rather conceptual, as the Bayesian update generally assumes $c_{ij,t} = 1$. Still, there may occur reasons for observation penalization ($c_{ij,t} < 1$), e.g., if an observation is far from the credible region of the predictive distribution. Such outlying observations may degrade the estimation. Thus, moderation of their impact by lower weights or complete rejection with $c_{ij,t} = 0$ may be necessary. This broad stand-alone theory is, however, beyond the scope of this paper; the reader is referred to relevant literature, e.g., [53], [54].

B. Combination weights $a_{ij,t}$

Our approach to the determination of the combiner weights $a_{ij,t}$ is consistent with the rest of the paper in that it is model-oriented. It exploits the idea that the node i prefers the neighbor in $\mathcal{I}_i$ whose estimates are more likely to be the best fit for the observations. This means, that if we arranged the weights $a_{ij,t}$ into a vector $a_{i,t}$, it has a categorical distribution parameterized by a probability vector $q_{i,t}$ of the same length as $a_{i,t}$. The probability mass function and expected values are

$$p_i(a_{i,t}|q_{i,t}) = \prod_{j\in\mathcal{I}_i} q_{ij,t}^{a_{ij,t}}, \quad \text{and} \quad \mathbb{E}\left[a_{ij,t}|q_{i,t}\right] = q_{ij,t}, \tag{18}$$

respectively. Hence the prediction of $a_{ij,t}$ relies on the knowledge of $q_{ij,t}$. Its Bayesian estimation is analytically tractable with the conjugate Dirichlet prior distribution with the hyperparameters $\psi_{ij,t-1}$ arranged into a vector $\psi_{i,t-1}$ aggregating the knowledge about $q_{i,t}$, and the probability density and expectations according to

$$\pi_i(q_{i,t}|\psi_{i,t-1}) = \frac{1}{B(\psi_{i,t-1})} \prod_{j\in\mathcal{I}_i} q_{ij,t}^{\psi_{ij,t-1}-1}, \tag{19}$$

$$\mathbb{E}[q_{ij,t}|\psi_{i,t-1}] = \frac{\psi_{ij,t-1}}{\sum_{j\in\mathcal{I}_i} \psi_{ij,t-1}}, \tag{20}$$

where $B(\cdot)$ is the multivariate beta function. Conceptually, if $a_{ij,t}$ were observable, the Bayesian update of the prior distribution (19) by observations from (18) would be

$$\begin{aligned} \pi_i(q_{i,t}|\psi_{i,t}) &\propto \prod_{j\in\mathcal{I}_i} q_{ij}^{a_{ij,t}} \prod_{j\in\mathcal{I}_i} q_{ij}^{\psi_{ij,t-1}-1} \\ &= \prod_{j\in\mathcal{I}_i} q_{ij}^{a_{ij,t}+\psi_{ij,t-1}-1}. \end{aligned} \tag{21}$$

In reality, it is not observable which node from $\mathcal{I}_i$ has the best knowledge of θ, but it is still possible to measure how well the nodes fit the observations using the predictive performance,

$$\mathcal{A}_{ij,t} = \int \pi_j(\theta|\zeta_{j,t-1}) \prod_{k\in\mathcal{I}_i} f_k(y_{k,t}|z_{k,t},\theta)\mathrm{d}\theta. \tag{22}$$

The quasi-Bayesian approach (similar to [55]) then advocates the use of the measure $\mathcal{A}_{ij,t}$ in place of $a_{ij,t}$ in (21). The expectations of $a_{ij,t}$ then follow from (18) and (20), respectively.

Remark 1. *The probabilistically consistent evaluation of the predictive distribution* (22) *via the chain rule and marginalization may be both analytically intractable and computationally demanding. In this case, it is possible to resort to a simple* plug-in principle*, using the posterior point estimate of θ directly in the observations-generating model. Naturally, this (asymptotically equivalent) procedure ignores uncertainty about θ and may significantly influence results under small sample sizes [56].*

Remark 2. *One interpretation of the combiner weights is that they are* probabilities *of correctly explaining the observed data by the neighbors' posterior distributions. This provides a way towards communication savings in a sense somewhat similar to [24], which in one extreme is achieved by removal of links [57]. At time t, node i samples a subset of nodes from $\mathcal{I}_i$ (of a fixed or random size) with probabilities determined by the Dirichlet distribution. Marginalization then provides the combiner weights. This procedure is probabilistically well founded and does not require additional communication steps.*

Algorithm 1 BAYESIAN ATC DIFFUSION ESTIMATION

```
The nodes $i = 1,\dots,I$ are initialized with the prior densities $\pi_i(\theta|\cdot)$. The Dirichlet prior hyperparameters $\psi_{i,0}$ for combination weights are set. For $t = 1,2,\dots$ and each node $i$ do:
Adaptation phase:
  1) Get observations $y_{j,t}$ and $z_{j,t}$ of neighbors $j \in \mathcal{I}_i$.
  2) Update the Dirichlet prior hyperparameters $\psi_{i,t}$, Equation (21) using the predictive density (22).
  3) Update the prior distribution of $\theta$, Equation (5), under conjugacy (6).
Combination phase:
  1) Calculate the point estimates $\mathbb{E}[a_{ij}|\cdot]$, Equation (18).
  2) Get the posterior densities $\pi_j(\theta|\cdot)$ of neighbors $j \in \mathcal{I}_i$.
  3) Combine the posterior densities according to Proposition 1, under conjugacy Equation (8).
```

VI. Example 1: Recursive linear regression

This example demonstrates the diffusion estimation of normal linear regression models with unknown noise variance. First, the Bayesian estimation is derived in terms of sufficient statistics and relevant hyperparameters. Part of these derivations can be found, e.g., in [58]. Then, the adaptation and combination phases are a straightforward application of the principles from Section III. The example also demonstrates the bijective mapping between the "standard" hyperparameters and their conjugate counterparts. A reader familiar with the inference under more complicated conjugate prior distributions may also notice one more quality: rewriting the densities and the subsequent inference in terms of ξ and ν is algebraically easier than deriving the posterior distributions under standard parameters.

Assume the linear regression model

$$y_t = \beta^\intercal z_t + \varepsilon_t,$$

where $t = 1,2,\dots$ is a discrete time index, y_t is a real scalar observation, $z_t \in \mathbb{R}^p$ is a regression vector, and $\beta \in \mathbb{R}^p$ is a vector of unknown constant regression coefficients. The i.i.d. univariate normal noise variables ε_t have zero mean and an unknown constant variance σ^2. The probability density of $y_t|z_t,\beta,\sigma^2$ is thus normal, and with a slight abuse of notation (see paragraph below Definition 1),

$$\begin{aligned} f(y_t|z_t,\beta,\sigma^2) &= \frac{(\sigma^2)^{-\frac{1}{2}}}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(y_t - \beta^\intercal z_t)^2\right\} \\ &= \frac{(\sigma^2)^{-\frac{1}{2}}}{\sqrt{2\pi}} \exp\left\{ \mathrm{Tr}\left(\underbrace{-\frac{1}{2\sigma^2}\begin{bmatrix}1\\-\beta\end{bmatrix}\begin{bmatrix}1\\-\beta\end{bmatrix}^\intercal}_{\eta} \underbrace{\begin{bmatrix}y_t\\z_t\end{bmatrix}\begin{bmatrix}y_t\\z_t\end{bmatrix}^\intercal}_{T(y_t,z_t)} \right)\right\}. \end{aligned}$$

The goal is to estimate $\theta = \{\beta,\sigma^2\}$ sequentially from the incoming observations. A convenient prior distribution, which is conjugate to the model, is the normal inverse-gamma distribution

$$\begin{aligned} \beta,\sigma^2 &\sim \mathcal{N}i\mathcal{G}(m_{t-1},V_{t-1},a_{t-1},b_{t-1}) \\ &= \mathcal{N}(m_{t-1},\sigma^2 V_{t-1}) \times \mathcal{G}(a_{t-1},b_{t-1}), \end{aligned}$$

with scalar positive hyperparameters a_{t-1} and b_{t-1}, a mean vector $m_{t-1} \in \mathbb{R}^p$ and a scaling matrix V_{t-1}^{-1} of appropriate dimensions. One can show that after a little algebra its density has the form

$$\pi(\beta, \sigma^2|\cdot) = \frac{b^{a_{t-1}}(\sigma^2)^{-(a_{t-1}+1+\frac{p}{2})}}{\sqrt{2\pi}|V_{t-1}|^{\frac{1}{2}}\Gamma(a_{t-1})} \exp\Bigg\{ -\frac{1}{2\sigma^2}\Bigg[b_{t-1}+ \\ + \mathrm{Tr}\left(\begin{bmatrix}1\\-\beta\end{bmatrix}^\intercal \begin{bmatrix}1\\-\beta\end{bmatrix} \begin{bmatrix}m_{t-1}^\intercal\\ I\end{bmatrix} V_{t-1}^{-1} \begin{bmatrix}m_{t-1}^\intercal\\ I\end{bmatrix}^\intercal\right)\Bigg]\Bigg\}.$$

The density reveals that the prior conjugate hyperparameters are given by

$$\xi_{t-1} = \begin{bmatrix} m_{t-1}^\intercal V_{t-1}^{-1} m_{t-1} + 2b_{t-1} & m_{t-1}^\intercal V_{t-1}^{-1} \\ V_{t-1}^{-1} m_{t-1} & V_{t-1}^{-1} \end{bmatrix} \\ = \begin{bmatrix} \xi_{t-1}^{[11]} & \xi_{t-1}^{[12]} \\ \xi_{t-1}^{[21]} & \xi_{t-1}^{[22]} \end{bmatrix},$$

(the latter block-matrix form will be used below) and

$$\nu_{t-1} = 2a_{t-1}.$$

From the Bayesian update (3), which is based on conjugate priors and models of the exponential family, we can easily derive the update of the original hyperparameters. We have

$$\begin{aligned} V_t &= \left(V_{t-1}^{-1} + z_t z_t^\intercal\right)^{-1} = V_{t-1} - \frac{V_{t-1} z_t z_t^\intercal V_{t-1}}{1 + z_t^\intercal V_{t-1} z_t} = \left(\xi_t^{[22]}\right)^{-1}, \\ m_t &= V_t(V_{t-1}^{-1} m_{t-1} + y_t z_t) = \left(\xi_t^{[22]}\right)^{-1} \xi_t^{[21]}, \\ a_t &= a_{t-1} + \frac{1}{2} = \frac{1}{2}(\nu_{t-1} + 1) = \frac{1}{2}\nu_t, \qquad (23) \\ b_t &= b_{t-1} + \frac{1}{2}\left(-m_t^\intercal V_t^{-1} m_t + m_{t-1}^\intercal V_{t-1}^{-1} m_{t-1} + y_t^2\right) \\ &= \frac{1}{2}\left[\xi_t^{[11]} - \xi_t^{[12]}\left(\xi_t^{[22]}\right)^{-1}\left(\xi_t^{[12]}\right)^\intercal\right], \end{aligned}$$

where the second equality for V_t follows from the Sherman-Morrison rank-one update formula. It is worth remarking that this is the point where the determination of the posterior hyperparameters V_t, m_t, and b_t from ξ_t is particularly easy.

It is immediately obvious that the marginal posterior distribution of σ^2 is $i\mathcal{G}(a_t, b_t)$ with mean and variance given by

$$\mathbb{E}[\sigma^2|\cdot] = \frac{b_t}{a_t - 1}, \quad \text{and} \quad \mathrm{var}(\sigma^2|\cdot) = \frac{b_t^2}{(a_t-1)^2(a_t-2)}.$$

Furthermore, it can be shown that the marginal posterior distribution of β is the generalized multivariate Student's t distribution with $2a_t$ degrees of freedom, location m_t, and scale matrix $\frac{b_t}{a_t} V_t$ [58]. Finally, the predictive distribution given z',

$$f(y'|y_{0:t}, z_{0:t}, z') = \int f(y'|z', \beta, \sigma^2)\pi(\beta, \sigma|y_{0:t}, z_{0:t})\mathrm{d}\beta\mathrm{d}\sigma^2$$

is the generalized multivariate Student's t distribution

$$y'|y_{0:t}, z_{0:t}, z' \sim t_{2a_t}\left(m_t^\intercal z', \frac{b_t}{a_t}\left(1 + (z')^\intercal V_t z'\right)\right).$$

A. Diffusion estimation

a) Adaptation phase: The diffusion adaptation phase – Equation (6) – in terms of the sufficient statistics $T(y_{j,t}, z_{j,t})$ and hyperparameters $\xi_{i,t-1}$ and $\nu_{i,t-1}$ is a straightforward application of Equation (6),

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

The original posterior hyperparameters $V_{i,t}, m_{i,t}, a_{i,t}$ and $b_{i,t}$ can be derived from $\xi_{i,t}$ as in (23).

b) Combination phase: The diffusion combination is a direct application of Proposition 1. Likewise, as in the diffusion adaptation, the original hyperparameters can be derived from the resulting $\widetilde{\xi}_{i,t}$ and $\widetilde{\nu}_{i,t}$.

Remark 3. *The rank-one update of $V_{i,t}$ can be easily implemented sequentially. First, assign $V_{i,t} \leftarrow V_{i,t-1}$. Then, for all $j \in \mathcal{I}_i$, do*

$$V_{i,t} \leftarrow V_{i,t} - \frac{c_{ij,t} V_{i,t} z_{j,t} z_{j,t}^\intercal V_{i,t}}{1 + c_{ij,t} z_{j,t}^\intercal V_{i,t} z_{j,t}}.$$

This equation, together with the update formula for $m_{i,t}$ coincide with the diffusion RLS (diffRLS) adaptation step of Cattivelli and Sayed [12]. There are two crucial differences between diffRLS and the proposed algorithm. First, diffRLS imposes the cumbersome requirement of known noise variances. If this knowledge is available, the prior reduces to the normal distribution and the adaptation steps fully coincide. The second difference lies in the combination phase. diffRLS combines only the point estimates, and leaves their covariances intact, which indeed saves communication resources, but may potentially degrade the estimation. The proposed algorithm combines whole posterior distributions consistently within the framework of probability and information theory.

VII. Example 2: GLMs & Logistic regression

This example demonstrates approximate estimation of generalized linear models (GLMs) in diffusion networks. The class of GLMs comprises several popular models including the linear regression, logistic, probit, and multinomial regression models. Below, we consider only scalar GLMs for simplicity, and focus on the logistic regression model.

Remark 4. *We remark that [33] recently proposed a diffusion stochastic optimization algorithm allowing estimation of certain GLMs too. The algorithm stems from the non-Bayesian paradigm, assumes unknown data distribution and non-smoothly regularized expectations of the loss function in the role of a risk function.*

Scalar GLMs are given by

$$\mathbb{E}\left[y_t|z_t, \theta\right] = g^{-1}(\theta^\intercal z_t),$$

where y_t is a scalar observation, $z_t \in \mathbb{R}^n$ an observable regressor, $\theta \in \mathbb{R}^n$ is a vector of unknown regression coefficients, and $g(\cdot)$ is a link function, whose argument $\theta^\intercal z_t$ is called the linear predictor (hence the name generalized *linear* models). The observations are i.i.d. and have an exponential family distribution. For example, the linear regression model arises if g is the identity function and y_t follows the normal distribution. The logistic regression model is characterized by dichotomous $y_t \in \{0, 1\}$ such that

$$y_t \sim Bernoulli(p_t),$$

where the parameter $p_t \in [0, 1]$ is the probability of $y_t = 1$. The probability mass function of y_t can be written as

$$f(y_t|p_t) = f(y_t|z_t, \theta) = p_t^{y_t}(1 - p_t)^{1-y_t}. \tag{24}$$

The logit link function defined on p_t is related to the linear predictor $\theta^\intercal z_t$ as follows:

$$g(p_t) = \operatorname{logit}(p_t) = \log\left(\frac{p_t}{1 - p_t}\right) = \theta^\intercal z_t,$$

and thus,

$$\mathbb{E}[y_t|z_t, \theta] = p_t = \operatorname{logit}^{-1}(\theta^\intercal z_t) = \frac{1}{1 + \exp(-\theta^\intercal z_t)}. \tag{25}$$

Direct Bayesian estimation of GLMs is often analytically intractable due to the lack of conjugate priors, but it is possible to resort to approximations by normal distributions. The posterior distribution is approximated by a normal distribution centered at the posterior mode and with the covariance equal to minus the inverse of the second derivative of the log posterior density at this mode. The accuracy is reasonable even under small-sample cases, as long as the approximated posterior is smooth and unimodal, or multimodal with a dominant mode [59]. More specifically,

$$\pi(\theta|y_t, z_t, \zeta_{t-1}) \propto \underbrace{f(y_t|z_t, \theta)}_{Eq.(24)} \underbrace{\pi(\theta|\zeta_{t-1})}_{\mathcal{N}(\hat{\theta}_{t-1}, \Sigma_{t-1})}$$

is approximated by $\tilde{\pi}(\theta|y_t, z_t, \zeta_{t-1})$ in two steps. First, the mode $\hat{\theta}_t$ is found, e.g., using Newton's iterative method. This step, thus, coincides with the maximum a posteriori (MAP) estimation. Second, the posterior covariance is calculated,

$$-\left[\frac{\partial^2 \log \pi(\theta|y_t, z_t, \zeta_{t-1})}{\partial\theta\partial\theta^\intercal}\right]^{-1}_{\theta=\hat{\theta}_t} = \left[\Sigma_{t-1}^{-1} + \hat{y}_t(1 - \hat{y}_t) z_t z_t^\intercal\right]^{-1},$$

where $\hat{y}_t = \mathbb{E}[y_t|z_t, \theta]$. The resulting approximating normal posterior density

$$\tilde{\pi}(\theta|y_t, z_t, \zeta_{t-1}) \approx \mathcal{N}\left(\hat{\theta}_t, \left[\Sigma_{t-1}^{-1} + \hat{y}_t(1 - \hat{y}_t) z_t z_t^\intercal\right]^{-1}\right) \tag{26}$$

is asymptotically approaching the true posterior density according to the Bayesian central limit theorem [48], [49].

The problem of tractability affects also the predictive distribution

$$f(y'|z', \zeta_t) = \int f(y'|z', \zeta_t, \theta)\pi(\theta|\zeta_t)\mathrm{d}\theta.$$

Again, there are methods for its approximation, mostly variations of the Laplace's method [59]–[61]. The most basic one yields [60]

$$\begin{aligned} f(y'|z', \zeta_t) \approx (2\pi)^{\frac{n}{2}} &\left(\det[\Sigma_t^{-1} + y'(1 - y')z'(z')^\intercal]\right)^{-\frac{1}{2}} \\ &\times f(y'|z', \theta)\tilde{\pi}(\theta|\zeta_t), \end{aligned}$$

where $\theta = \hat{\theta}_t$ is used.

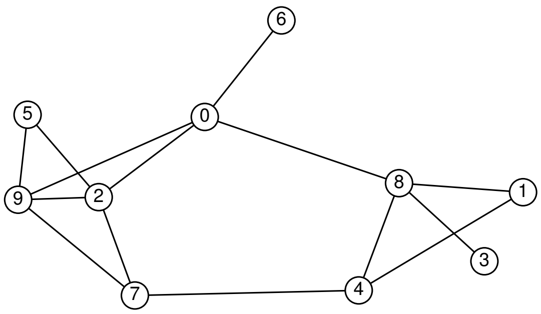


Fig. 3. The network layout used in the logistic regression experiments.

A. Diffusion estimation

a) Adaptation phase: The diffusion adaptation phase is based on Equation (5) and the above-described approximation of the true posterior density by a normal density (26). The logarithm of the true density has the form

$$\begin{aligned} &\log \pi_i(\theta|\zeta_{i,t}) = \log \pi_i(\theta|\zeta_{i,t-1}) \\ &+ \sum_{j \in \mathcal{I}_i} \log c_{ij,t}\left[\left(\frac{1}{1 + e^{-\theta^\intercal z_{j,t}}}\right)^{y_{j,t}}\left(1 - \frac{1}{1 + e^{-\theta^\intercal z_{j,t}}}\right)^{1-y_{j,t}}\right]. \end{aligned}$$

Plugging its maximum, i.e., the mode $\hat{\theta}_{i,t}$, into its negative inverse differential yields the covariance of the approximating normal density (26),

$$\Sigma_{i,t} = \left[\Sigma_{i,t-1}^{-1} + \sum_{j \in \mathcal{I}_i} c_{ij,t}\hat{y}_{j,t}(1 - \hat{y}_{j,t}) z_{j,t} z_{j,t}^\intercal\right]^{-1},$$

where $\hat{y}_{j,t} = [1 + exp(-\hat{\theta}_{i,t}^\intercal z_{j,t})]^{-1}$ is the point prediction calculated at node i using the neighbors' regressors, c.f. (25).

b) Combination phase: The combination phase is a straightforward application of Proposition 1, specifically Lemma 3.

B. Numerical Example

The numerical examples demonstrate the performance of four methods: the ATC method, the adaptation-only method (denoted A), the combination-only method (denoted C), and the noncooperative scenario (denoted NOCOOP), where the network nodes do not collaborate at all. The network, depicted in Figure 3, consisted of 10 nodes. The regression vectors $z_{i,t} \in \mathbb{R}^4$ had a '1' as a first entry (an offset term) and random samples from $\mathcal{U}(-1, 1)$ for the remaining entries. The elements of $\theta \in \mathbb{R}^4$ were independently sampled from $\mathcal{U}(-2, 2)$.

The initial normal prior distribution identical for all the network nodes had a zero mean vector and a diagonal covariance matrix $100 I_{4\times 4}$. The weights $c_{ij,t}$ were determined adaptively. However, we stress that the homogeneous conditions allow for taking advantage of uniform weights that remove the computation of likelihoods while yielding practically identical results. The results were averaged over 200 experiment runs.

Figure 4 depicts the evolution of MSD averaged over the network. It is not surprising that the collaboration improves the estimation quality, particularly if the posterior distributions are shared (as discussed in Section IV-D).

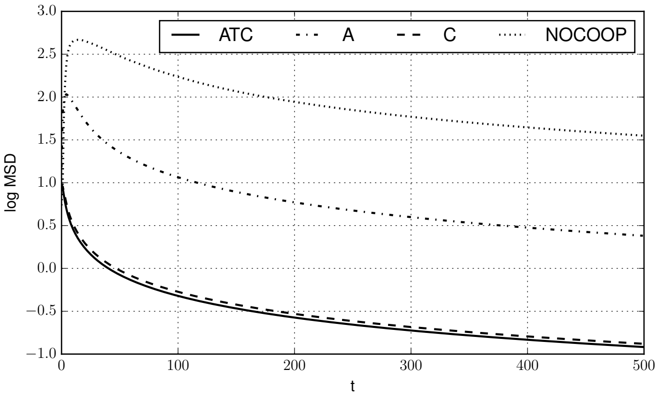


Fig. 4. Evolution of the decimal logarithm of MSD averaged over the network.

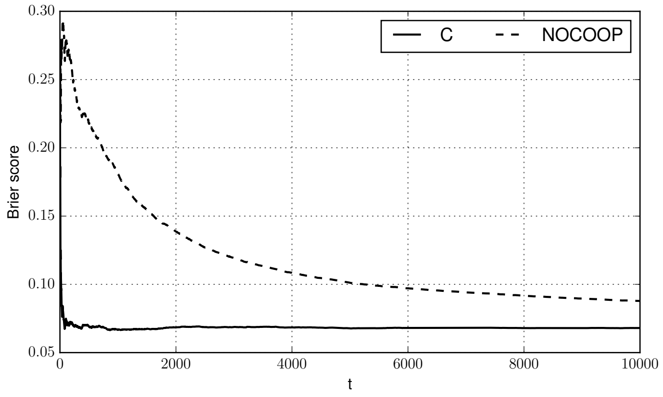


Fig. 5. Network Brier score of the combination-only and non-collaborative modes.

C. Example: skin–non-skin classification

The second example considers the skin–non-skin dataset of Bhatt and Dhall [62]. It consists of 245,057 samples of which 50,859 are skin samples and 194,198 are non-skin samples. The dataset was collected by randomly sampling RGB values from face images of various age groups (young, middle, and old), race groups (white, black, and Asian), and gender. The samples were obtained from the FERET and PAL databases. Each data item consisted of four variables – B, G, R and the class label.

Our goal was to estimate the logistic model parameters where the regression vectors were $z_{i,t} = [1, B_{i,t}, G_{i,t}, R_{i,t}]^{\intercal}$ (the first term standing for the offset) and the dependent variable $y_{i,t}$ denotes the class (0 is skin, 1 is non-skin). The data were randomly shuffled before processing and were introduced sequentially. The normal prior for θ was the same as in the previous example.

The network of 10 nodes was the same as in the previous example. Each node sequentially acquired 10,000 observations. Two scenarios were compared — the combination-only mode with uniform weights (C) and the noncooperative mode (NOCOOP). This allowed for fair determination of the predictive ability of the methods based on their prediction of skin class membership $\mathbb{E}[y_{i,t}|z_{i,t}, \theta]$. The decision cutoff value was 0.5.

A popular categorical classifier assessment measure is the Brier score [63], expressing the predictor performance as follows:

$$\mathrm{B}_i = \frac{1}{T}\sum_{t=1}^{T}(y_{i,t} - \mathbb{E}[y_{i,t}|z_{i,t}, \theta])^2.$$

The Brier score is equivalent to MSE. Its average calculated over the network is depicted in Figure 5. Apparently, collaboration led to very fast stabilization of predictions. This is also confirmed by the sensitivity (true positive rate), specificity (true negative rate), accuracy and the diagnostic odds ratio (prevalence-independent accuracy, DOR) at $t = 1,500$ and $t = 10,000$, Table II. A centralized strategy — merging of posterior distributions in a fusion center — led to virtually the same performance, with a negligibly lower diagnostic odds ratio (78.520) caused by a few additional misclassifications.

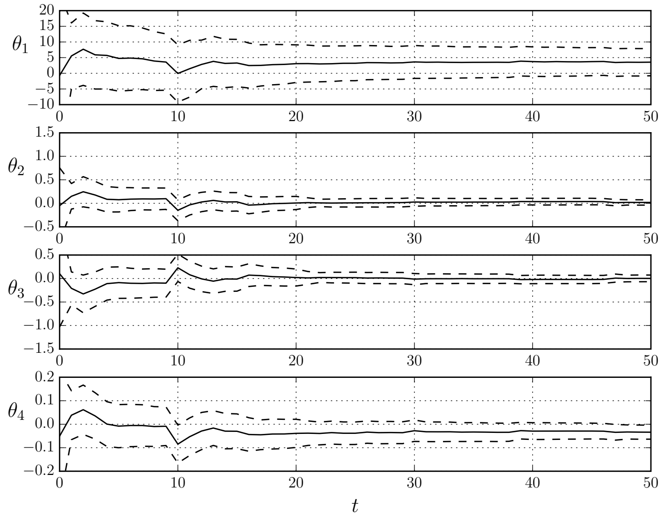


Fig. 6. Evolution of estimates with ±3 standard deviation bands (node 0, first 50 time steps).

The maximum likelihood estimate of θ obtained from 10,000 samples was $\hat{\theta} = [4.611, 0.031, -0.014, -0.034]^{\intercal}$. The estimates of all the elements were statistically significant. Figure 6 shows that under collaboration the Bayesian estimates very quickly converged to these values.

VIII. Example 3: Kalman filtering

This example considers estimation of state-space models by Kalman filtering. The diffusion Kalman filter is derived for the linear case, but it is natural that the extended and other types of Kalman filters follow the same principles. In the sequel, the filtering is cast in terms of the conjugate hyperparameter ξ so that its diffusion variant is easily obtained. For the sake of completeness, the update of the original hyperparameters is shown as well.

Let us assume a state-space model of the form

$$x_t|x_{t-1}, z_t \sim \mathcal{N}\left(A_t x_{t-1} + B_t z_t, Q_t\right), \qquad (27)$$

$$y_t|x_t \sim \mathcal{N}\left(H_t x_t, R_t\right), \qquad (28)$$

where $y_t \in \mathbb{R}^k$ is an observation at time $t = 1, 2, \ldots$, $x_t \in \mathbb{R}^n$ is a state vector, $z_t \in \mathbb{R}^n$ is an observable input variable,

TABLE II
SENSITIVITY (TRUE POSITIVE RATE), SPECIFICITY (TRUE NEGATIVE RATE), ACCURACY, AND DIAGNOSTIC ODDS RATIO (DOR) AT $t = 1,500$ AND $t = 10,000$ FOR THE COMBINATION-ONLY (C) AND NONCOLLABORATIVE (NOCOOP) MODES.

Measure	$t = 1,500$		$t = 10,000$	
	C	*NOCOOP*	*C*	*NOCOOP*
Sensitivity	0.829	0.689	0.823	0.775
Specificity	0.944	0.875	0.945	0.935
Accuracy	0.920	0.837	0.919	0.902
DOR	81.723	15.508	78.891	49.547

A_t, B_t and H_t are matrices of compatible dimensions, and $Q_t \in \mathbb{R}^{n \times n}$ and $R_t \in \mathbb{R}^{k \times k}$ are state and observation covariance matrices.

The Bayesian sequential inference of the state vector x_t from past observations is based on a Gaussian prior probability density function $\pi(x_t|y_{0:t-1}, z_{0:t-1})$ whose mean and covariance matrix are denoted by x_t^- and P_t^-. After incorporating y_t and z_t, we obtain the filtering distribution $\pi(x_t|y_{0:t}, z_{0:t})$, which is also Gaussian and with mean and covariance x_t^+ and P_t^+, respectively. The Kalman filtering proceeds in two steps:

a) Prediction: the predicted value of x_t from x_{t-1} is obtained using the state evolution model (27) and the Chapman-Kolmogorov equation,

$$\pi(x_t|y_{0:t-1}, z_{0:t}) = \int \pi(x_t|x_{t-1}, z_t) \pi(x_{t-1}|y_{0:t-1}, z_{0:t-1}) dx_{t-1}.$$

The properties of normal distributions ensure that the predicted prior probability density function $\pi(x_t|y_{0:t-1}, z_{0:t})$ is again a normal distribution $\mathcal{N}(x_t^-, P_t^-)$ with hyperparameters

$$x_t^- = A_t x_{t-1}^+ + B_t z_t, \quad \text{and} \quad P_t^- = A_t P_{t-1}^+ A_t^\intercal + Q_t.$$

In diffusion networks, the prediction step is run locally without any collaboration.

b) Update: The Bayes' theorem (1) serves for updating the prior of x_t with the information about x_t in the observed y_t and z_t,

$$\pi(x_t|y_{0:t}, z_{0:t}) = \frac{\pi(x_t|y_{0:t-1}, z_{0:t}) f(y_t|x_t)}{\int \pi(x_t|y_{0:t-1}, z_{0:t}) f(y_t|x_t) dx_t}. \tag{29}$$

As before, we rewrite the observation model (28) in the exponential family form,

$$\begin{aligned} f(y_t|x_t) &\propto \exp\left\{ -\frac{1}{2}(y_t - H_t x_t)^\intercal R_t^{-1} (y_t - H_t x_t) \right\} \\ &= \exp\left\{ \mathrm{Tr}\left(\underbrace{-\frac{1}{2}\begin{bmatrix} -1 \\ x_t \end{bmatrix}\begin{bmatrix} -1 \\ x_t \end{bmatrix}^\intercal}_{\eta} \underbrace{\begin{bmatrix} y_t^\intercal \\ H_t^\intercal \end{bmatrix} R_t^{-1} \begin{bmatrix} y_t^\intercal \\ H_t^\intercal \end{bmatrix}^\intercal}_{T(y_t)} \right) \right\}. \end{aligned}$$

The conjugate normal distribution in the corresponding compatible form is given by

$$\begin{aligned} \pi(x_t|y_{0:t-1}, z_{0:t}) &\propto \exp\left\{ -\frac{1}{2}(x_t - x_t^-)^\intercal (P_t^-)^{-1} (x_t - x_t^-) \right\} \\ &= \exp\left\{ \mathrm{Tr}\left(\underbrace{-\frac{1}{2}\begin{bmatrix} -1 \\ x_t \end{bmatrix}\begin{bmatrix} -1 \\ x_t \end{bmatrix}^\intercal}_{\eta} \underbrace{\begin{bmatrix} (x_t^-)^\intercal \\ I \end{bmatrix} (P_t^-)^{-1} \begin{bmatrix} (x_t^-)^\intercal \\ I \end{bmatrix}^\intercal}_{\xi_t} \right) \right\}, \end{aligned}$$

where I is a unit matrix of appropriate size.

The Bayesian update (29) then reduces to the update of the hyperparameters according to Equation (3),

$$\begin{aligned} \xi_t &= \xi_{t-1} + T(y_t) \\ &= \begin{bmatrix} (x_t^-)^\intercal (P_t^-)^{-1} x_t^- + y_t^\intercal R_t^{-1} y_t, & (x_t^-)^\intercal (P_t^-)^{-1} + y_t^\intercal R_t^{-1} H_t \\ (P_t^-)^{-1} (x_t^-)^\intercal + H_t^\intercal R_t^{-1} y_t, & (P_t^-)^{-1} + H_t^\intercal R_t^{-1} H_t. \end{bmatrix} \end{aligned}$$

Now, it is easy to derive the diffusion estimator.

A. Diffusion estimation

a) Adaptation phase: A direct application of Equation (6) shows that

$$\begin{aligned} \xi_{i,t} &= \xi_{i,t-1} + \sum_{j \in \mathcal{I}_i} c_{ij} \begin{bmatrix} y_{j,t} \\ H_{j,t} \end{bmatrix} R_{j,t}^{-1} \begin{bmatrix} y_{j,t} \\ H_{j,t} \end{bmatrix}^\intercal, \\ \nu_{i,t} &= \nu_{i,t-1} + 1, \end{aligned} \tag{30}$$

from which it easily follows that

$$\begin{aligned} P_{i,t}^+ &= \left[(P_{i,t}^-)^{-1} + \left(\sum_{j \in \mathcal{I}_i} c_{ij} H_{j,t}^\intercal R_{j,t}^{-1} H_{j,t} \right) \right]^{-1}, \\ x_{i,t}^+ &= x_{i,t}^- + P_{i,t}^+ \left[\sum_{j \in \mathcal{I}_i} c_{ij} H_{j,t}^\intercal R_{j,t}^{-1} \left(y_{j,t} - H_{j,t} x_{i,t}^- \right) \right]. \end{aligned}$$

b) Combination phase: Proposition 1 applied to $\xi_{j,t}$ from Equation (30) is straightforward. Since the posterior densities are normal, it is already known from Lemma 3 that the original hyperparameters — the mean and covariance matrix — take the form

$$\begin{aligned} \widetilde{P}_{i,t}^+ &= \left[\sum_{j \in \mathcal{I}_i} a_{ij,t} \left(P_{j,t}^+ \right)^{-1} \right]^{-1}, \\ \widetilde{x}_{i,t}^+ &= \widetilde{P}_{i,t}^+ \left(\sum_{j \in \mathcal{I}_i} a_{ij,t} \left(P_{j,t}^+ \right)^{-1} x_{j,t}^+ \right). \end{aligned}$$

Remark 5. *The original diffusion Kalman filter (diffKF) is due to Cattivelli and Sayed [31]. Both algorithms have the same adaptation phase but different combination phase, as diffKF combines only local state estimates and leaves the associated covariances intact. Hu, Xie and Zhang [32] extend diffKF by a covariance intersection-based merging, yielding an algorithm equivalent to the one proposed in this paper. Therefore, a numerical example is omitted here. Instead, Figure 7 illustrates the risk associated with the negligence of the (co)variance properties.*

IX. EXAMPLE 4: INHOMOGENEOUS POISSON PROCESS

This example demonstrates the diffusion estimation of a slowly varying parameter with a scheduled combination phase. Its purpose is twofold. The first is to demonstrate that the application of certain Bayesian techniques is straightforward and does not require separate development. A simple exponential forgetting procedure (scaling of the prior distribution) illustrates this. The second purpose is to shed some light on further research possibilities.

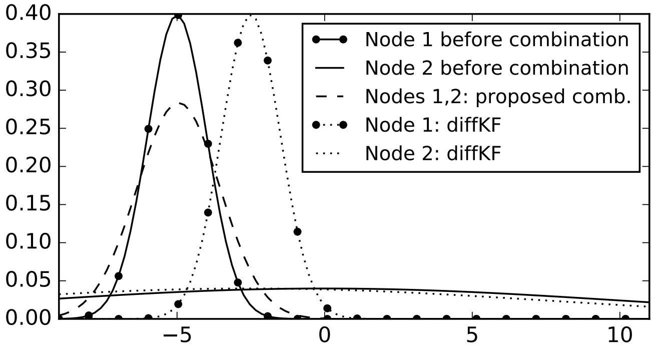


Fig. 7. A simple example of combination methods. 2 nodes share their distributions $\mathcal{N}(-5,1)$ and $\mathcal{N}(0,100)$ (e.g. a freshly connected node), respectively (solid lines), the combination weights are uniform. The proposed combination procedure yields in both nodes $\mathcal{N}(-4.95,1.98)$ (dashed line), while the combination of point estimates only (as in diffKF) leads to $\mathcal{N}(-2.5,1)$ and $\mathcal{N}(-2.5,100)$ (dotted lines).

The Poisson process, also known as shot noise process, finds applications in many disciplines including astronomy, physics, image processing, and telecommunications. It characterizes stochastically independent events like the number of particles colliding with a detector, the number of phone calls, Internet traffic and many other phenomena that occur in a given interval [64]. It is a memoryless discrete random process with independent increments, describing the number $y_{(t_a,t_b]}$ of events that occur between two time instances t_a and t_b,

$$y_{(t_a,t_b]} \sim \mathcal{P}o\left(\theta_t\left(t_b - t_a\right)\right), \qquad t_b \geq t_a,$$

where the probability mass function of $y_{(t_a,t_b]}$ has the form

$$f(y_{(t_a,t_b]}|\theta_t) = \frac{[\theta_t(t_b-t_a)]^{y_{(t_a,t_b]}}}{y_{(t_a,t_b]}!} e^{-\theta_t(t_b-t_a)}, \;\; y_{(t_a,t_b]} \in \mathbb{N}_0.$$

The parameter θ_t is real and positive and represents the process intensity. If the intensity is time-varying, the process is called inhomogeneous. In the sequel, we will consider sampling of $y_{(t_a,t_b]}$ in regular time intervals of unit length, i.e., $t_b = t_a + 1$, and denote the observations (number of occurrences) between two consecutive time instants by y_t.

Under process homogeneity, a convenient conjugate prior distribution is the gamma distribution $\mathcal{G}(\xi_{t-1}, \nu_{t-1})$ with scalar hyperparameters $\xi_{t-1}, \nu_{t-1} > 0$ and density

$$\pi(\theta|y_{0:t-1}) = \frac{\nu_{t-1}^{\xi_{t-1}}}{\Gamma(\xi_{t-1})} \theta^{\xi_{t-1}-1} e^{-\nu_{t-1}\theta}.$$

It is straightforward to see that the posterior hyperparameters obey Equations (3) with $T(y_t) = y_t$. The posterior mean and variance are

$$\mathbb{E}[\theta|y_{0:t}] = \frac{\xi_t}{\nu_t} \quad \text{and} \quad \mathrm{var}(\theta|y_{0:t}) = \frac{\xi_t}{\nu_t^2}.$$

Inhomogeneity of the Poisson process impairs the estimation tractability. However, under mild variations of θ_t, a way of circumventing the intractability is by way of using the concept of forgetting. Forgetting amounts to flattening of the posterior distribution before incorporating new observations. In order to avoid distraction from the main subject of this paper, we adopt only the most basic exponential forgetting [58]. The forgetting factor λ is a positive real number, where $0 < \lambda \leq 1$, and it is usually close to one. The factor is used to flatten the prior distribution by exponentiating it, i.e., $\widetilde{\pi}(\theta|y_{0:t-1}) \leftarrow [\pi(\theta|y_{0:t-1})]^{\lambda}$.

It is easy to show that the predictive density has the form

$$f(y'_{t+s}|\bar{y}_{0:t}) = \frac{s^{y'_{t+s}}}{y'_{t+s}!} \frac{\nu_t^{\xi_t}}{(\nu_t+s)^{\xi_t+y'_{t+s}}} \frac{\Gamma(\xi_t + y'_{t+s})}{\Gamma(\xi_t)}.$$

A. Diffusion estimation

a) Adaptation phase: The diffusion update (5) with forgetting takes the form

$$\pi_i(\theta_t|\bar{y}_{0:t}) \propto [\pi_i(\theta_t|\bar{y}_{0:t-1})]^{\lambda} \prod_{j\in\mathcal{I}_i} [f(y_{j,t}|\theta_t)]^{c_{ij}},$$

and in terms of the gamma hyperparameters,

$$\xi_{i,t} = \lambda\xi_{i,t-1} + \sum_{j\in\mathcal{I}_i} c_{ij} y_{j,t}, \quad \text{and} \quad \nu_{i,t} = \lambda\nu_{i,t-1} + \sum_{j\in\mathcal{I}_i} c_{ij}.$$

b) Combination phase: As usually, the combination phase follows Proposition 1. Unlike in the previous examples, ξ and ν are “standard” parameters of the gamma distribution, and therefore, there is no reparameterization.

B. Numerical Example

This example demonstrates the estimation (tracking) of slowly varying parameter θ_t by a network of 20 nodes depicted in Figure 8. The datasets of 500 observations were randomly generated from the Poisson distribution with $\theta_t = 5 + 2\sin(500\pi/t)$. The prior gamma distribution was initialized with hyperparameters $\nu_{i,0} = \xi_{i,0} = 0.1$, and the exponential forgetting factor was set to $\lambda = 0.96$. The investigated scenarios were ATC with adaptive combiners, adaptation-only (A), combination-only with uniform combiners (C), noncooperative scenario (NOCOOP), and a combination-only scenario with uniform combiners and a scheduled combination phase activated at every 5th time step (C5). The last scheme suggests one possible way towards communication savings. The resulting MSD evolutions that are depicted in Figure 9 were averaged over 200 experiments. The ATC strategy, again, exhibited the best performance, and it was closely followed by the combination-only scenario.

The trade-off between the estimation performance and the communication costs can be very easily tuned by the scheduled combination phase. The figure also displays the ‘wavy’ character of the MSD caused by the impact of outdated information that is being gradually forgotten. A typical evolution of estimates of ATC and NOCOOP of a randomly chosen node is shown in Figure 10, where the impact of past data is clearly visible on the much smoother ATC estimates.

X. Conclusion and further remarks

In this paper we proposed a Bayesian approach to sequential diffusion estimation. The main objective was to present an abstract methodology that is straightforwardly applicable to inference of parameters of a wide class of popular models.

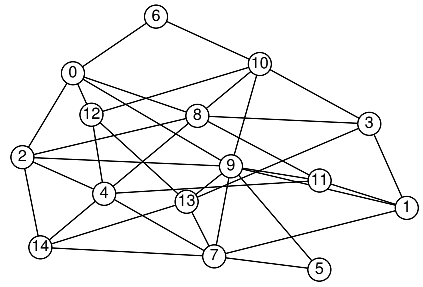


Fig. 8. Poisson process estimation: Network layout.

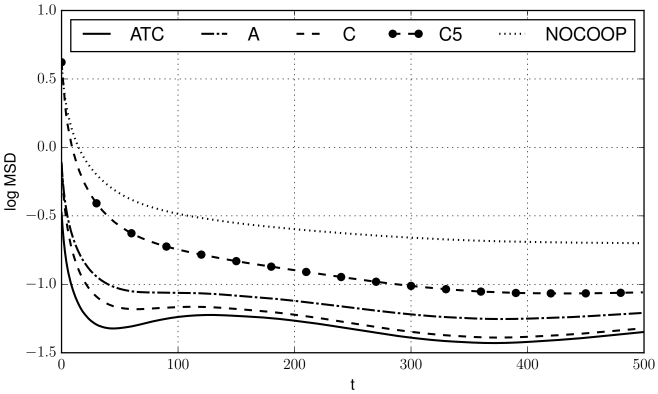


Fig. 9. Evolution of the decimal logarithm of MSD averaged over the network. The character of the curves is caused by the delayed influence of forgetting, c.f. Fig. 10. Still the variations are within a well acceptable interval.

The methodology was demonstrated on four examples. The recursive linear regression and the diffusion Kalman filter already have their counterparts in the state-of-art literature (though our approach to regression does not necessarily assume known noise variances). The diffusion logistic regression and the diffusion tracking of (inhomogeneous) Poisson process intensity are new.

Probably the most important aspect of the adopted Bayesian viewpoint is its straightforward application. For instance, the last example illustrates tracking of slowly varying parameter using a forgetting technique. Similarly, it is possible to approximate unimodal posterior distributions obtained from a sequential Monte Carlo filter by an exponential family distribution, combine them according to Proposition 1, and resample from the result. The framework is also applicable to sequential mixtures estimation whose basics were proposed [36]. All this suggests that the proposed framework opens up many possible directions for research.

The source codes can be found at *http://diffest.utia.cas.cz.*

REFERENCES

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, “A survey on sensor networks,” *IEEE Communications Magazine*, vol. 40, no. 8, pp. 102–114, Aug. 2002.

[2] Y. Liu, C. Li, W. Tang, and Z. Zhang, “Distributed estimation over complex networks,” *Information Sciences*, vol. 197, pp. 91–104, Aug. 2012.

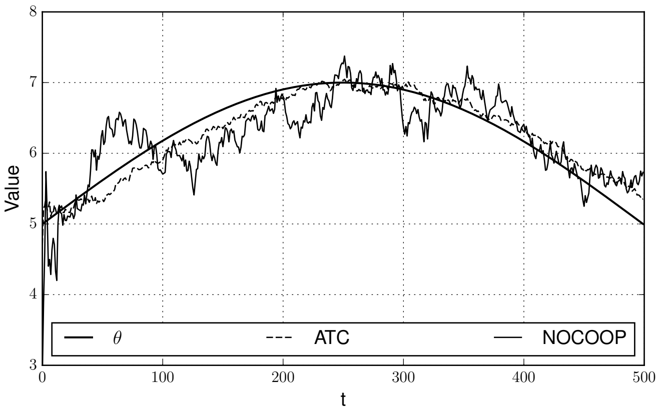


Fig. 10. A typical evolution of estimates of θ for two marginal cases, ATC and the noncollaborative scenario (NOCOOP). The influence of fading past observations is apparent from the delayed character of the ATC estimates. The noncollaborative scenario provides more scattered estimates due to the insufficient information (less observations).

[3] A. H. Sayed, “Adaptation, learning, and optimization over networks,” *Foundations and Trends in Machine Learning*, vol. 7, no. 4-5, pp. 311–801, 2014.

[4] ——, “Adaptive networks,” *Proc. IEEE*, vol. 102, no. 4, pp. 460–497, Apr. 2014.

[5] J. N. Tsitsiklis and M. Athans, “Convergence and asymptotic agreement in distributed decision problems,” in *Proc. 21st IEEE Conference on Decision and Control*, pp. 692–701, Dec. 1982.

[6] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, “Distributed asynchronous deterministic and stochastic gradient optimization algorithms,” in *Proc. American Control Conference 1984*, Jun. 1984, pp. 484–489.

[7] D. P. Bertsekas, “A new class of incremental gradient methods for least squares problems,” *SIAM Journal on Optimization*, vol. 7, no. 4, pp. 913–926, 1997.

[8] A. Nedić and D. P. Bertsekas, “Incremental subgradient methods for nondifferentiable optimization,” *SIAM Journal on Optimization*, vol. 12, no. 1, pp. 109–138, 2001.

[9] M. Rabbat and R. Nowak, “Quantized incremental algorithms for distributed optimization,” *IEEE J. Sel. Areas Commun.*, vol. 23, no. 4, pp. 798–808, Apr. 2005.

[10] C. G. Lopes and A. H. Sayed, “Incremental adaptive strategies over distributed networks,” *IEEE Trans. Signal Process.*, vol. 55, no. 8, pp. 4064–4077, Aug. 2007.

[11] J. Plata-Chaves, N. Bogdanović, and K. Berberidis, “Distributed incremental-based RLS for node-specific parameter estimation over adaptive networks,” in *Proc. 21st European Signal Processing Conference (EUSIPCO)*, pp. 1–5, 2013.

[12] 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.

[13] 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.

[14] A. Bertrand, M. Moonen, and A. H. Sayed, “Diffusion bias-compensated RLS estimation over adaptive networks,” *IEEE Trans. Signal Process.*, vol. 59, no. 11, pp. 5212–5224, 2011.

[15] X. Zhao, S.-Y. Tu, and A. H. Sayed, “Diffusion adaptation over networks under imperfect information exchange and non-stationary data,” *IEEE Trans. Signal Process.*, vol. 60, no. 7, pp. 3460–3475, Jul. 2012.

[16] K. Dedecius and V. Sečkárová, “Dynamic diffusion estimation in exponential family models,” *IEEE Signal Process. Lett.*, vol. 20, no. 11, pp. 1114–1117, Nov. 2013.

[17] R. Arablouei, K. Dogancay, S. Werner, and Y.-F. Huang, “Adaptive distributed estimation based on recursive least-squares and partial diffusion,” *IEEE Trans. Signal Process.*, vol. 62, no. 14, pp. 3510–3522, Jul. 2014.

[18] M. H. DeGroot, “Reaching a consensus,” *Journal of the American Statistical Association*, vol. 69, no. 345, pp. 118–121, Mar. 1974.

[19] R. Olfati-Saber and R. M. Murray, "Consensus problems in networks of agents with switching topology and time-delays," *IEEE Trans. Autom. Contr.*, vol. 49, no. 9, pp. 1520–1533, Sep. 2004.

[20] 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.

[21] I. D. Schizas, G. B. Giannakis, and Z.-Q. Luo, "Distributed estimation using reduced-dimensionality sensor observations," *IEEE Trans. Signal Process.*, vol. 55, no. 8, pp. 4284–4299, Aug. 2007.

[22] I. D. Schizas, G. Mateos, and G. B. Giannakis, "Distributed LMS for consensus-based in-network adaptive processing," *IEEE Trans. Signal Process.*, vol. 57, no. 6, pp. 2365–2382, Jun. 2009.

[23] M. B. Guldogan, "Consensus Bernoulli filter for distributed detection and tracking using multi-static Doppler shifts," *IEEE Signal Process. Lett.*, vol. 21, no. 6, pp. 671–675, Jun. 2014.

[24] M. G. S. Bruno and S. S. Dias, "Collaborative emitter tracking using Rao-Blackwellized random exchange diffusion particle filtering," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 1, p. 19, Feb. 2014.

[25] P. Braca, S. Marano, and V. Matta, "Running consensus in wireless sensor networks," in *Proc. 11th International Conference on Information Fusion*, Jun. 2008, pp. 1–6.

[26] S. Kar and J. M. F. Moura, "Gossip and distributed Kalman filtering: Weak consensus under weak detectability," IEEE Trans. Signal Process., vol. 59, no. 4, pp. 1766–1784, Apr. 2011.

[27] 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.

[28] P. Di Lorenzo and A. H. Sayed, "Sparse distributed learning based on diffusion adaptation," *IEEE Trans. Signal Process.*, vol. 61, no. 6, pp. 1419–1433, Mar. 2013.

[29] J. Chen, C. Richard, A. O. Hero, and A. H. Sayed, "Diffusion LMS for multitask problems with overlapping hypothesis subspaces," in *Proc. 2014 IEEE Workshop on Machine Learning for Signal Processing (MLSP2014)*, Reims, France, 2014.

[30] J. Plata-Chaves, N. Bogdanović, and K. Berberidis, "Distributed diffusion-based LMS for node-specific adaptive parameter estimation," *IEEE Trans. Signal Process.*, vol. 63, no. 13, pp. 3448–3460, Jul. 2015.

[31] 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, 2010.

[32] J. Hu, L. Xie, and C. Zhang, "Diffusion Kalman filtering based on covariance intersection," *IEEE Trans. Signal Process.*, vol. 60, no. 2, pp. 891–902, Feb. 2012.

[33] S. Vlaski, L. Vandenberghe, and A. H. Sayed, "Diffusion stochastic optimization with non-smooth regularizers," in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 4149–4153.

[34] Z. J. Towfic, J. Chen, and A. H. Sayed, "Collaborative learning of mixture models using diffusion adaptation," in *Proc. 2011 IEEE International Workshop on Machine Learning for Signal Processing*. IEEE, Sep. 2011, pp. 1–6.

[35] S. S. Pereira, R. Lopez-Valcarce, and A. Pages-Zamora, "A diffusion-based EM algorithm for distributed estimation in unreliable sensor networks," *IEEE Signal Process. Lett.*, vol. 20, no. 6, pp. 595–598, Jun. 2013.

[36] K. Dedecius, J. Reichl, and P. M. Djurić "Sequential estimation of mixtures in diffusion networks," *IEEE Signal Process. Lett.*, vol. 22, no. 2, 197–2001, 2014.

[37] O. Hlinka, F. Hlawatsch, and P. M. Djurić "Distributed particle filtering in agent networks: A survey, classification, and comparison," *Signal Processing Magazine, IEEE*, vol. 30(1), pp. 61–81, 2013.

[38] K. Dedecius and P. M. Djurić, "Diffusion filtration with approximate Bayesian computation," in *Proc. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Apr. 2015, pp. 3207–3211.

[39] A. Raftery, M. Kárný, and P. Ettler, "Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill." *Technometrics*, vol. 52, no. 1, pp. 52–66, Feb. 2010.

[40] K. Dedecius, "Diffusion estimation of state-space models: Bayesian formulation," in *Proc. 2014 IEEE Workshop on Machine Learning for Signal Processing (MLSP2014)*. Sep. 2014, pp. 1–6.

[41] K. Dedecius and V. Sečkárová, "Distributed modelling of big dynamic data with generalized linear models," in *Proc. 17th Intl. Conf. on Information Fusion*, 2014.

[42] C. P. Robert, *The Bayesian Choice*. Springer, Jun. 2007.

[43] B. O. Koopman, "On distributions admitting a sufficient statistic," *Trans. Am. Math. Soc.*, vol. 39, no. 3, pp. 399–409, May 1936.

[44] H. Raiffa and R. Schlaifer, *Applied Statistical Decision Theory*. Harvard University Press, Jan. 1961.

[45] G. Battistelli, L. Chisci, C. Fantacci, A. Farina, and A. Graziano, "Consensus CPHD filter for distributed multitarget tracking," *IEEE J. Sel. Top. Signal Process.*, vol. 7, no. 3, pp. 508–520, Jun. 2013.

[46] L. Chisci, G. Battistelli, C. Fantacci, A. Farina, A. Graziano, and R. Mahler, "Distributed fusion of multitarget densities and consensus PHD filters," in *Proc. 18th Intl. Conf. on Information Fusion*, 2015.

[47] E. L. Lehmann and G. Casella, *Theory of Point Estimation*. Springer, Aug. 1998.

[48] J. M. Bernardo and A. F. M. Smith, *Bayesian theory*, John Wiley & Sons, Ltd., 2001.

[49] I. A. Ibragimov and R. Z. Hasminskii, "Asymptotic behavior of some statistical estimators II. Limit theorems for the a posteriori density and Bayes' estimators," *Theory of Probability & Its Applications*, vol. 18, no. 1, pp. 76–91, 1973.

[50] Y. Wang and P. M. Djurić, "Distributed Bayesian Estimation of Linear Models With Unknown Observation Covariances, *IEEE Transactions on Signal Processing*, vol. 64, no. 8, pp. 1962–1971, 2016.

[51] O. Bunke and X. Milhaud, "Asymptotic behavior of Bayes estimates under possibly incorrect models," Ann. Stat., vol. 26, no. 2, pp. 617–644, Apr. 1998.

[52] T. Choi and R. V. Ramamoorthi, "Remarks on consistency of posterior distributions," in Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh, IMS, vol. 3, pp. 170–186, 2008.

[53] P. Congdon, *Applied Bayesian modelling*. Wiley, 2003.

[54] C. C. Aggarwal, *Outlier Analysis*. Springer, 2013.

[55] M. Kárný, J. Kadlec, and E. L. Sutanto, "Quasi-Bayes estimation applied to normal mixture," in *Proc. 3rd European IEEE Workshop on Computer-Intensive Methods in Control and Data Processing*. 1998, pp. 77–82.

[56] R. L. Smith, "Bayesian and frequentist approaches to parametric predictive inference," *Bayesian Statistics*, vol. 6, pp. 589–612, 1999.

[57] X. Zhao and A. H. Sayed, "Single-link diffusion strategies over adaptive networks," in *Proc. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Mar. 2012, pp. 3749–3752.

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

[59] L. Tierney and J. B. Kadane, "Accurate approximations for posterior moments and marginal densities," *Journal of the American Statistical Association*, vol. 81, no. 393, pp. 82–86, Mar. 1986.

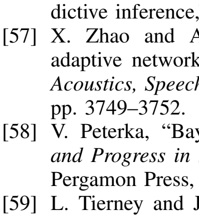
[60] A. Raftery, "Approximate Bayes factors and accounting for model uncertainty in generalised linear models," *Biometrika*, vol. 83, no. 2, pp. 251–266, Jun. 1996.

[61] M. Chen and X. Wang, "Approximate predictive densities and their applications in generalized linear models," *Computational Statistics & Data Analysis*, vol. 55, no. 4, pp. 1570–1580, 2011.

[62] R. B. Bhatt and A. Dhall, "Skin Segmentation Dataset." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Skin+Segmentation

[63] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, Jan. 1950.

[64] D. R. Insua, F. Ruggeri, and M. P. Wiper, *Bayesian Analysis of Stochastic Process Models*. Wiley, 2012.



Kamil Dedecius received the Ph.D. degree in Engineering Informatics from the Czech Technical University in Prague, Czech Republic, in 2010. Since 2010, he has been a Postdoc and a Research Assistant with the Institute of Information Theory and Automation, Czech Academy of Sciences. His primary research interests include mainly Bayesian probability and statistics, in particular the estimation theory and its application in signal processing. Since 2013 he focuses on the theory of fully distributed estimation in diffusion networks. His work has been recognized by the 2015 Otto Wichterle Award.

Petar M. Djurić (M'90–SM'99–F'06) received the B.S. and M.S. degrees in electrical engineering from the University of Belgrade, Belgrade, in 1981 and 1986, respectively, and the Ph.D. degree in electrical engineering from the University of Rhode Island, Kingston, RI, in 1990. Since 1990, he has been a Professor with the Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY. From 1981 to 1986, he was a Research Associate with the Vinča Institute of Nuclear Sciences, Belgrade. His research interests include the area of signal and information processing with primary interests in the theory of signal modeling, detection, and estimation; Monte Carlo-based methods; signal and information processing over networks; and applications in a wide range of disciplines. He has been invited to lecture at many universities in the United States and overseas. He received the IEEE Signal Processing Magazine Best Paper Award in 2007 and the EURASIP Technical Achievement Award in 2012. In 2008, he was the Chair of Excellence of Universidad Carlos III de Madrid-Banco de Santander. From 2008 to 2009, he was a Distinguished Lecturer of the IEEE Signal Processing Society. He has been on numerous committees of the IEEE Signal Processing Society and of many professional conferences and workshops. He is a Fellow of EURASIP and the Editor-in- Chief of the IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS.