BAYESIAN ESTIMATION OF UNKNOWN PARAMETERS OVER NETWORKS

Petar M. Djurić[*]

Dept. of Electrical & Computer Engineering
Stony Brook University
Stony Brook, NY 11794, USA
e-mail: petar.djuric@stonybrook.edu

Kamil Dedecius [†]

Inst. of Information Theory and Automation,
Czech Academy of Sciences
Pod Vodárenskou věží 1143/4
182 08 Prague 8, Czech Republic
e-mail: dedecius@utia.cas.cz

ABSTRACT

We address the problem of sequential parameter estimation over networks using the Bayesian methodology. Each node sequentially acquires independent observations, where all the observations in the network contain signal(s) with unknown parameters. The nodes aim at obtaining accurate estimates of the unknown parameters and to that end, they collaborate with their neighbors. They communicate to the neighbors their latest posterior distributions of the unknown parameters. The nodes fuse the received information by using mixtures with weights proportional to the predictive distributions obtained from the respective node posteriors. Then they update the fused posterior using the next acquired observation, and the process repeats. We demonstrate the performance of the proposed approach with computer simulations and confirm its validity.

Index Terms— parameter estimation over networks, Bayes theory, mixture models, model averaging

1. INTRODUCTION

In networks of interconnected nodes, an important class of inference problems is the signal and information processing with local collaboration and where no central unit exists. The nodes typically acquire their own observations, and they process them in order to detect, estimate, or classify the observations. They share with their neighbors their findings in terms of point estimates [1], decisions [2], or probability distributions [3], and use them to improve their inference. These types of problems are of growing interest, and their solutions find applications in diverse areas including, social sciences, engineering, and biology.

An important issue in problems of inference over networks is how the shared information is exploited optimally by the nodes for improved learning. In this paper, we are interested in exploring ways of doing it by relying on Bayesian theory. More specifically, we assume that the nodes acquire observations that contain signals with the same functional forms. Furthermore, the nodes know that some, if not all, of the unknown signal parameters of interest have the same values in the observations of the different nodes. The signals are distorted by independent noise processes that may have different strengths for different nodes or may be even of different type. Our interest is in sequential processing of the data, where at a given time instant t a node uses its prior to update it to a posterior that contains the information in the observation received at t. The node then shares this posterior with its neighbors. From the received posteriors, the nodes create a prior of the unknown parameters that will be used for processing the next observation. Thus, we assume that there is only one exchange of information among the neighboring nodes between two received observations.

A key operation for improving the learning from neighbors is the fusion of all the posteriors of the neighbors. In this paper, we propose that we linearly combine these posteriors, thus creating a mixture distribution. We also propose that the mixing coefficients of this mixture are proportional to the respective predictive distributions of the observations evaluated at new observations and where the predictive distributions are based on the nodes' posteriors. One can readily show that the mixing coefficients evolve with time as the accuracy of prediction of the posteriors increases or decreases. It is important to point out that since the overall posterior is a mixture distribution, one has to make sure that the size of these mixtures does not grow with time. This entails that before sharing the posteriors with the neighbors, the nodes have to reduce the mixture complexity.

In the next section, we present the problem. In Section 3, we explain the proposed procedure for learning from neighbors. In the following Section 4, we briefly outline ways for approximating the posterior distributions that are shared among neighbors. In Section 5, we present simulations of a normal linear regression process that demonstrate the performance of the proposed approach. Finally, in Section 6 we provide our concluding remarks.

[*]The first author was supported by NSF under Award CCF-1320626.

[†]The second author was supported by the Czech Science Foundation, grant No. 14-06678P.

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

2. PROBLEM FORMULATION

Suppose that we have a connected network with N nodes that only communicate with their neighbors. We denote the set of indices of neighbors of node i by $\mathcal{N}_i$. The index of node i is also in $\mathcal{N}_i$. Each node receives private observations $y_{i,t} \in \mathbb{R}^{d_y}$ sequentially in time, where $y_{i,t}$ is the observation of the ith node at t with $t \in \mathbb{N}^0$, and d_y is the dimension of $y_{i,t}$. The observations of the nodes are modeled by a parametric distribution given by $\ell(y_{i,t}|\theta)$, where $\theta \in \mathbb{R}^{d_\theta}$ is a parameter vector that is unknown to all the nodes, and d_θ is the dimension of θ.

We discriminate among several distributions. They are

$p(\theta\|\mathcal{I}_{i,t-1})$:	prior of θ at t,
$\pi(\theta\|y_{i,t},\mathcal{I}_{i,t-1})$:	posterior of θ at t,
$\widetilde{\pi}(\theta\|y_{i,t},\mathcal{I}_{i,t-1})$:	approximated posterior of θ at t,
$\ell(y_{i,t}\|\theta)$:	distribution of $y_{i,t}$ given θ,
$p(y_{i,t}\|y_{j,t-1},\mathcal{I}_{i,t-2})$:	predictive distribution of $y_{i,t}$.

In the above list, $\mathcal{I}_{i,t}$ symbolizes the information acquired by node i *by* time t and which is quantified by the posterior distributions of all the nodes in $\mathcal{N}_i$, or formally

$$\mathcal{I}_{i,t}: \quad \{\widetilde{\pi}(\theta|y_{j,\tau},\mathcal{I}_{j,\tau-1}), j \in \mathcal{N}_i, \tau = 0, 1, \cdots, t\}. \quad (1)$$

It goes without saying that in $\mathcal{I}_{i,t}$

$$\widetilde{\pi}(\theta|y_{i,\tau},\mathcal{I}_{i,\tau-1}) \quad = \quad \pi(\theta|y_{i,\tau},\mathcal{I}_{i,\tau-1}), \quad (2)$$

that is, the own posteriors of node i are not approximated. For succinct notation, in the fusion of information of node i, we will symbolize $\pi(\theta|y_{i,\tau},\mathcal{I}_{i,\tau-1})$ by $\widetilde{\pi}(\theta|y_{i,\tau},\mathcal{I}_{i,\tau-1})$.

We assume that at $t = 0$ each node has an initial distribution denoted by $\pi(\theta|\mathcal{I}_{i,-1})$, where $\mathcal{I}_{i,-1}$ is all the information available to the node at the beginning of modeling. These distributions are approximated (if necessary) by $\widetilde{\pi}(\theta|\mathcal{I}_{i,-1})$ and broadcasted to the neighbors. After receiving all the initial distributions from the neighbors, each node forms a fused prior of θ. The new prior of node i is denoted by $p(\theta|\mathcal{I}_{i,0})$, and we express the formation of the prior by

$$\{\widetilde{\pi}(\theta|\mathcal{I}_{j,-1})\}_{j\in\mathcal{N}_i} \quad \Longrightarrow \quad p(\theta|\mathcal{I}_{i,0}). \quad (3)$$

After receiving the first observations, each node exploits Bayes' theory to update its knowledge about θ, i.e., it forms a posterior according to

$$\pi(\theta|y_{i,1},\mathcal{I}_{i,0}) \quad \propto \quad \ell(y_{i,1}|\theta)p(\theta|\mathcal{I}_{i,0}), \quad (4)$$

where $\propto$ stands for "proportional to."

Once the nodes obtain their posteriors according to (4), they approximate them with $\widetilde{\pi}(\theta|y_{j,1},\mathcal{I}_{j,0})$ for reasons that will become clear in the sequel. Then they broadcast the approximated posteriors to their neighbors. Every node that receives such posteriors aims again at fusing this information with the information in its own posterior by implementing

$$\{\widetilde{\pi}(\theta|y_{j,1},\mathcal{I}_{j,0})\}_{j\in\mathcal{N}_i} \quad \Longrightarrow \quad p(\theta|\mathcal{I}_{i,1}). \quad (5)$$

The distribution $p(\theta|\mathcal{I}_{i,1})$ is a prior of θ of node i for processing the next observation $y_{i,2}$.

Now, the next observation is received and the process is repeated. First, the posteriors $\pi(\theta|y_{j,2},\mathcal{I}_{j,1})$ are formed, they are approximated and then broadcasted to the neighbors. With these posteriors, the nodes form new priors $p(\theta|\mathcal{I}_2)$ and so on.

The main problem is formulated as follows: how should the nodes process all the available posterior distributions so that they will have improved learning of the unknown parameters θ?

3. PROPOSED LEARNING

We address the stated problem by relying on the Bayesian theory. We view the distributions received from the neighbors as different "models" for describing the observed data. In the Bayesian literature, when one deals with different models, typically one resorts to model selection. In model selection, the basic task is to select the best model from a set of considered models using a predefined criterion. In our problem, however, we will not aim at picking the best distribution and keeping it, but instead we will preserve all the available distributions through the concept of model averaging [4, 5].

In our setting, a node does not know anything about the quality of the data of the neighbors used for obtaining their posteriors. All a node knows is the received parameter posteriors from the neighbors. The node then wants to use all the available information to form one posterior. The main question then is how to fuse the available posteriors. We propose to use as a posterior a mixture of all the posteriors where the weights of the mixands of the posterior are sequentially updated. This formation of posteriors will be performed in a principled way. Next we describe the main idea in more detail.

The focus will be on node i. At time $t = 0$, this node receives the initial prior distributions of the parameter θ, $\widetilde{\pi}(\theta|\mathcal{I}_{j,-1})$, from all its neighbors, i.e., $j \in \mathcal{N}_i$. The node i assigns weights to each of the neighbors (including itself) denoted by $w_{j,0}$, where $w_{j,0} \geq 0$, and $\sum_{j\in\mathcal{N}_j} w_{j,0} = 1$. Then the node forms its prior by

$$p(\theta|\mathcal{I}_{i,0}) \quad = \quad \sum_{j\in\mathcal{N}_i} w_{j,0}\, \widetilde{\pi}(\theta|\mathcal{I}_{j,-1}). \quad (6)$$

If the node has no prior history of "trust" in its neighbors, it assigns equal weights, i.e., $w_{j,0} = 1/|\mathcal{N}_i|$, with $|\mathcal{N}_i|$ representing the size of the set $\mathcal{N}_i$.

Next, the node i receives the first observation and evaluates the posterior of θ by

$$\begin{aligned} \pi(\theta|y_{i,1},\mathcal{I}_{i,0}) \quad &\propto \quad \ell(y_{i,1}|\theta)p(\theta|\mathcal{I}_{i,0}) \\ &= \quad \ell(y_{i,1}|\theta) \sum_{j\in\mathcal{N}_i} w_{j,0}\, \widetilde{\pi}(\theta|\mathcal{I}_{j,-1}). \end{aligned} \quad (7)$$

Thus, the posterior of θ of node i is a mixture distribution with $|\mathcal{N}_i|$ components. It is clear that if the nodes report these distributions at the next time instant, the number of mixands will quickly explode and will become unmanageable. The only way to proceed from here on is to obtain an approximation of $\pi(\theta|y_{i,1},\mathcal{I}_{i,0})$, by $\widetilde{\pi}(\theta|y_{i,1},\mathcal{I}_{i,0})$. There are many approaches of finding an approximation $\widetilde{\pi}(\theta|y_{i,1},\mathcal{I}_{i,0})$, and this will be addressed further below. Here we only maintain that for $t \geq 1$, if the approximation $\widetilde{\pi}_{i,1}(\theta|y_{i,1})$ is a mixture, it always has the same number of components.

At time $t = 1$ node i also quantifies the "quality" of its own prior and the priors received by its neighbors. To that end, we propose that the node treats the priors as models for describing future data. In other words, we propose that the quality of the priors of the nodes is evaluated by the predictive distributions obtained by the priors.

First we write the predictive distribution of $y_{i,1}$,

$$\begin{aligned} p(y_{i,1}|\mathcal{I}_{i,0}) &= \int_\Theta \ell(y_{i,1}|\theta) \sum_{j\in\mathcal{N}_j} w_{j,0}\widetilde{\pi}(\theta|\mathcal{I}_{j,-1})\mathrm{d}\theta \\ &= \sum_{j\in\mathcal{N}_i} w_{j,0} \int_\Theta \ell(y_{i,1}|\theta)\,\widetilde{\pi}(\theta|\mathcal{I}_{j,-1})\mathrm{d}\theta \\ &= \sum_{j\in\mathcal{N}_j} w_{j,0} p(y_{i,1}|\mathcal{I}_{j,-1}), \end{aligned} \tag{8}$$

where

$$p(y_{i,1}|\mathcal{I}_{j,-1}) = \int_\Theta \ell(y_{i,1}|\theta)\widetilde{\pi}(\theta|\mathcal{I}_{j,-1})\mathrm{d}\theta \tag{9}$$

is the predictive distribution of $y_{i,1}$ by using the posterior of node $j \in \mathcal{N}_i$, $\widetilde{\pi}(\theta|\mathcal{I}_{j,-1})$, and where the predictive distribution is computed at $y_{i,1}$.

The quality of $\widetilde{\pi}(\theta|\mathcal{I}_{j,-1})$ is measured by the weight $w_{j,0}$. The trust in node j is now updated according to

$$\begin{aligned} w_{j,1} &\propto w_{j,0}\int_\Theta \ell(y_{i,1}|\theta)\widetilde{\pi}(\theta|\mathcal{I}_{j,-1})\mathrm{d}\theta \\ &= w_{j,0}p(y_{i,1}|\mathcal{I}_{j,-1}). \end{aligned} \tag{10}$$

We note that the weights $w_{j,1}$ will be used in the formation of the priors of θ while processing the observations $y_{i,2}$.

From the above, the sequential nature of the processing of the next observations is clear. First the nodes approximate their posteriors $\pi(\theta|y_{j,1},\mathcal{I}_{j,0})$ with $\widetilde{\pi}(\theta|y_{j,1},\mathcal{I}_{j,0})$ and transmit them to their neighbors. Before $y_{i,2}$ is observed, the node i forms the prior of θ obtained from its own posterior and the received posteriors of the neighbors $\widetilde{\pi}(\theta|y_{j,1},\mathcal{I}_{j,0})$ and according to

$$p(\theta|\mathcal{I}_{i,1}) = \sum_{j\in\mathcal{N}_i} w_{j,1}\,\widetilde{\pi}(\theta|y_{j,1},\mathcal{I}_{j,0}). \tag{11}$$

This prior is updated to the posterior $\pi(\theta|y_{i,2},\mathcal{I}_{i,1})$ by

$$\pi(\theta|y_{i,2},\mathcal{I}_{i,1}) \propto \ell(y_{i,2}|\theta)\; p(\theta|\mathcal{I}_{i,1}). \tag{12}$$

The distribution $\pi(\theta|y_{i,2},\mathcal{I}_{i,1})$ is approximated by a simpler distribution, $\widetilde{\pi}(\theta|y_{i,2},\mathcal{I}_{i,1})$, and it is transmitted to its neighbors. Then the new weights of the neighbors of i are computed by

$$w_{j,2} \propto w_{j,1}p(y_{i,2}|y_{j,1}\mathcal{I}_{j,0}), \tag{13}$$

where

$$p(y_{i,2}|y_{j,1},\mathcal{I}_{j,0}) = \int_\Theta \ell(y_{i,2}|\theta)\widetilde{\pi}(\theta|y_{j,1},\mathcal{I}_{j,0})\mathrm{d}\theta. \tag{14}$$

The process continues following the same steps.

The summary of the implementation of the method is described below. The description is given for node i. At time $t = 0$, the node assigns weights to its neighbors $w_{j,0}$ and it receives from the neighboring nodes the distributions $\widetilde{\pi}(\theta|\mathcal{I}_{j,-1})$. It is assumed that at time $t-1$ (where $t = 1, 2, \cdots$) the node i receives posteriors of θ from its neighbors, $\widetilde{\pi}(\theta|y_{j,t-1},\mathcal{I}_{j,t-2})$. Also, at that time the node has its updated weights $w_{j,t-1}$, $j \in \mathcal{N}_i$.

Operations performed at time t:

1. Construction of the fused prior,

$$p(\theta|\mathcal{I}_{i,t-1}) = \sum_{j\in\mathcal{N}_i} w_{j,t-1}\,\widetilde{\pi}(\theta|y_{j,t-1},\mathcal{I}_{j,t-2}).$$

2. Computation of the posterior,

$$\pi(\theta|y_{i,t},\mathcal{I}_{i,t-1}) \propto \ell(y_{i,t}|\theta)\; p(\theta|\mathcal{I}_{i,t-1}).$$

3. Approximation of the posterior $\pi(\theta|y_{i,t},\mathcal{I}_{i,t-1})$ by $\widetilde{\pi}(\theta|y_{i,t},\mathcal{I}_{i,t-1})$.

4. Transmission of $\widetilde{\pi}(\theta|y_{i,t},\mathcal{I}_{i,t-1})$ to the neighbors.

5. Update of the weights $w_{j,t-1}$ to $w_{j,t}$ by

$$\widetilde{w}_{j,t} = w_{j,t-1}p(y_{i,t}|y_{j,t-1},\mathcal{I}_{j,t-2}),$$

and

$$w_{j,t} = \frac{\widetilde{w}_{j,t}}{\sum_{k\in\mathcal{N}_j}\widetilde{w}_{k,t}},$$

and where

$$p(y_{i,t}|y_{j,t-1},\mathcal{I}_{j,t-2}) = \int_\Theta \ell(y_{i,t}|\theta)\widetilde{\pi}(\theta|y_{j,t-1},\mathcal{I}_{j,t-2})\mathrm{d}\theta.$$

4. APPROXIMATIONS OF THE POSTERIORS

There are many methods for approximating the mixtures with simpler distributions. Here we mention a few.

Perhaps the simplest approach is to use the moment matching method. Suppose we want to approximate $\pi(\theta|y_{i,t},\mathcal{I}_{i,t-1})$ with $\widetilde{\pi}(\theta|y_{i,t},\mathcal{I}_{i,t-1})$. We compute the

moments of θ with respect to $\pi_{i,t}(\theta|y_{i,t})$ and choose the parameters of $\widetilde{\pi}_{i,t}(\theta|y_{i,t})$ so that the moments of θ remain the same.

One big class of methods is based on exploiting a measure of divergence of two densities (e.g., the Kullback-Leibler divergence). The main idea is to choose a distribution $\widetilde{\pi}(\theta|y_{i,t}, \mathcal{I}_{i,t-1})$ that comes from a certain admissible family and that minimizes the dissimilarity of $\pi(\theta|y_{i,t}, \mathcal{I}_{i,t-1})$ and $\widetilde{\pi}(\theta|y_{i,t}, \mathcal{I}_{i,t-1})$. For example, one can aim at mixture reduction by way of merging the mixture components by using the Kullback-Leibler divergence [6].

Another set of approaches can be found in [7], where the Occam's window method is applied. In [8], an approach based on function approximation is proposed. The approximation minimizes an upper bound of the approximation error between the original and the simplified model as measured by the L^2 distance.

5. SIMULATIONS

The following simulation example demonstrates the properties of the proposed collaborative estimation method. We consider a network of six nodes $i = \{0, \dots, 5\}$ depicted in Fig. 1. Each node receives 400 observations generated by a normal regression process

$$y_{i,t} = \theta^\top x_{i,t} + \varepsilon_{i,t},$$

where $\theta = [0.4, -0.8, 0.3, 0.1]^\top$ is the common vector of regression coefficients, $x_{i,t} \sim \prod_{j=1}^{4} \mathcal{U}_j(0,1)$ are nodes-specific regressors, and the zero-mean normal noise $\varepsilon_{i,t}$ has a standard deviation 0.9 at node 3 and 0.2 elsewhere.

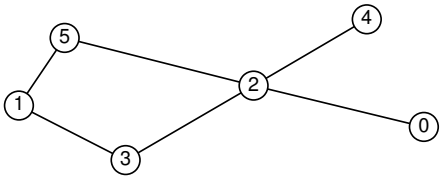

Fig. 1. Topology of the network used for simulation.

The estimation of θ exploits the multivariate normal prior distributions $\mathcal{N}(\mu_{i,t-1}, \Sigma_{i,t-1})$, initialized with $\mu_{i,0} = [0,0,0,0]^\top$ and $\Sigma_{i,0} = 100 \cdot I_{4\times 4}$. The prior weights are uniform. The fusion procedure is based on the moment matching method and yields a single normal distribution with mixture moments as parameters,

$$\widetilde{\mu}_{i,t-1} = \sum_{j\in\mathcal{N}_i} w_{j,t-1}\mu_{j,t-1},$$

$$\widetilde{\Sigma}_{i,t-1} = \sum_{j\in\mathcal{N}_i} w_{j,t-1}\left(\mu_{j,t-1}\mu_{j,t-1}^\top + \Sigma_{j,t-1}\right) - \widetilde{\mu}_{i,t-1}\widetilde{\mu}_{i,t-1}^\top.$$

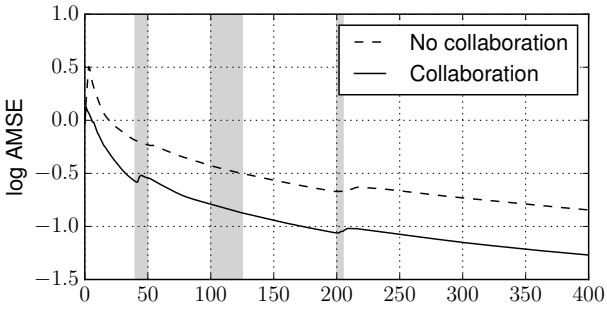

Fig. 2. Evolution of the decimal logarithm of the mean squared error averaged over the network. The three grey bands correspond to failure periods.

In order to demonstrate the robustness of the proposed method, three failure periods were introduced:

- $t \in \{40, \dots, 50\}$ — The prior distribution of node 4 was reset to the initial prior distribution.
- $t \in \{100, \dots, 125\}$ — Node 5 measurements were void, that is, $y_{5,t} = 0$ in this period.
- $t \in \{200, \dots, 205\}$ — The prior distribution of node 3 was reset to the initial prior distribution.

Otherwise, the process is stable. Due to this stability, the exchanged distributions may gradually get highly similar, which influences the dynamics of the weights. This effect is suppressed by increasing their uncertainty by means of exponential flattening, independently introduced in [9, 10], and used in a similar fashion in dynamic model averaging [11].

We compare two scenarios. First, the nodes collaborate according to the proposed method, i.e., they share their distributions with neighbors, and proceed with them with adaptively inferred weights. Second, the nodes do not collaborate at all, and perform the parameter inference only locally.

The estimation performance is measured by the mean squared error averaged over the network (AMSE). Its evolution is depicted in Fig. 2 for both scenarios. The collaboration improved the estimation quality considerably. The reactions to failures are fast and relatively stable. Figure 3 shows the weights evolutions at the network nodes. From the plots, we can conclude that the proposed algorithm has several appealing features. First of all, it effectively discounts nodes with a higher noise variance (Node 3). Second, it immediately reacts to failures and quickly recovers from them. Finally, the weights assigned to neighbors with similar statistical properties tend to concentrate under regular (failure-free) conditions.

6. CONCLUSIONS

In this paper we proposed the concept of model averaging for fusion of information in sequential estimation over networks. The nodes of a network process their own observations by using the Bayesian paradigm. They exchange their posteriors with their neighbors which are then fused to form a prior for processing the next observations. The obtained priors are mixtures with mixand coefficients that keep evolving with time and that reflect the predictive performance of the distributions received by the respective neighbors. The proposed method was tested on a network with six nodes where the nodes observe a regression process and with three types of disruptions. The results revealed several appealing features of the method.

7. REFERENCES

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

[2] P. M. Djurić and Y. Wang, "Distributed Bayesian learning in multiagent systems: Improving our understanding of its capabilities and limitations," *Signal Processing Magazine, IEEE*, vol. 29, no. 2, pp. 65–76, 2012.

[3] K. Dedecius, J. Reichl, and P. M. Djurić, "Sequential estimation of mixtures in diffusion networks.," *IEEE Signal Processing Letters*, vol. 22, no. 2, pp. 197–201, 2015.

[4] C. Gerda and N. L. Hjort, *Model Selection and Model Averaging*, vol. 330, Cambridge University Press Cambridge, 2008.

[5] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, "Bayesian model averaging: A tutorial," *Statistical science*, pp. 382–401, 1999.

[6] T. Ardeshiri, U. Orguner, and E. Özkan, "Gaussian mixture reduction using reverse Kullback-Leibler divergence," *arXiv preprint arXiv:1508.05514*, 2015.

[7] D. Madigan and A. E. Raftery, "Model selection and accounting for model uncertainty in graphical models using Occam's window," *Journal of the American Statistical Association*, vol. 89, no. 428, pp. 1535–1546, 1994.

[8] K. Zhang and J. T. Kwok, "Simplifying mixture models through function approximation," *Neural Networks, IEEE Transactions on*, vol. 21, no. 4, pp. 644–658, 2010.

[9] J.Q. Smith, "The multiparameter steady model," *Journal of Royal Statistical Society, Ser. B*, vol. 43, pp. 256–260, 1981.

[10] V. Peterka, "Bayesian approach to system identification," *Trends and Progress in System identification*, vol. 1, pp. 239–304, 1981.

[11] A.E. 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, 2010.

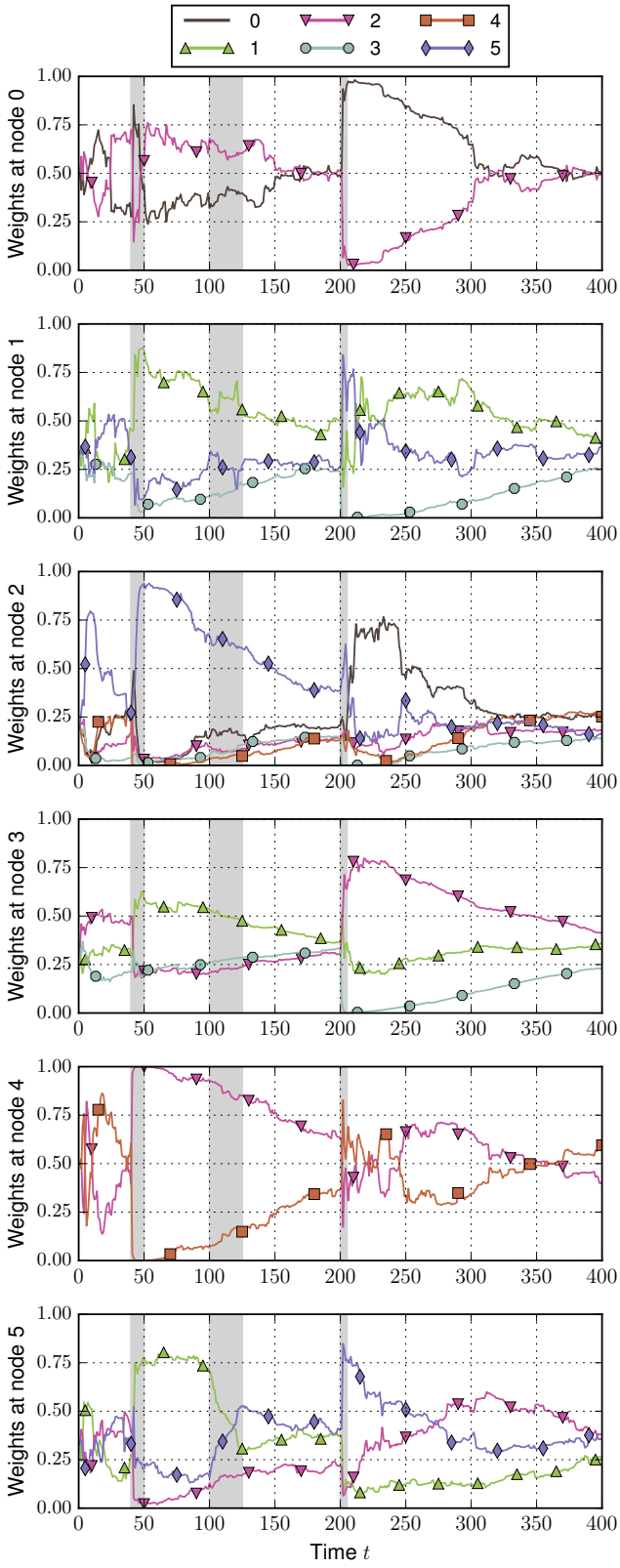

Fig. 3. Evolution of neighbors' weights at nodes 0 to 5. The three grey bands correspond to failure periods.