ITERATIVE LEARNING OF DFT-DOMAIN DYNAMICAL MODELS SUBJECT TO PARAMETER VARIATIONS

Sarmad Malik and Gerald Enzner

Institute of Communication Acoustics, Ruhr-Universität Bochum, 44780 Germany
phone: + (49) 234 32 25872, fax: + (49) 234 32 14165
email: {sarmad.malik,gerald.enzner}@rub.de
web: www.ruhr-uni-bochum.de/ika

ABSTRACT

We present a DFT-domain expectation-maximization framework for maximum-likelihood learning of linear dynamical models. The expectation step takes the form of a diagonalized DFT-domain Kalman filter coupled with a fixed-lag smoother, which effectively traces the evolution of the hidden state for a given underlying dynamical model defined via its covariance parameters. The maximization step learns the covariance parameters of the dynamical model and specifically discerns itself from a conventional algorithm by yielding distinct outputs for each block within the lag interval. Hence, in our approach the reliance on a fixed-lag for expressing the complete data likelihood does not necessarily entail the traditional conjecture of stationarity for the system within the duration of the lag interval. The capability to account for possible non-stationarity further helps the devised algorithm to carry out optimal and mutually synergetic state estimation and model inference, which we comprehensively substantiate with the help of simulation results.

1. INTRODUCTION

The advantages of describing real world systems by linear dynamical models has over the years attracted considerable attention towards analysis, estimation and learning of model parameters in applications pertaining to biomedical data analysis, automatic control, object tracking, etc., [1, 2]. Not only do dynamical models allow the explicit expression of *a priori* beliefs about the state and evolution of the unknown system, but they also provide a useful framework for including all important factors which influence the observation.

State estimators for dynamical models have been generally presented in the form of Kalman filter and its variants [3, 4], which consistently, and rightly so, rely heavily upon the parameters of the underlying model. The *parameters* of a dynamical model are unknown quantities, which the estimator assumes as known and lacks the capability to learn as such. Therefore, a state estimator can be flaunted to be an optimal one in a given statistical sense only if the parameters of the model are assumed known. The assumption of knowing model parameters is surely not a trivial one, which at times needs to be tediously justified via tuning or application specific knowledge. Thus, a natural progression in the analysis of dynamical models points towards parameter learning mechanisms [5]. In [1] the EM algorithm has been construed as a generic framework for joint estimation tasks, i.e., state estimation and parameter learning.

Motivated by the need for efficient signal processing, we consider a DFT-domain representation for block-wise processing of linear dynamical systems as proposed in [6]. Thus we express the complete data likelihood in DFT-domain, which is necessary to realize a maximum-likelihood expectation maximization (ML-EM) algorithm over a lag interval. We differentiate our approach from works in [5] and [7] by modeling block-time dependent model parameters within the lag period, which leads to recursive learning of parameters over the duration of the lag interval and improves parameter learning of the algorithm in non-stationary environments.

The complete data likelihood leads to a posterior which, unlike a traditional filtering distribution, assumes knowledge of a finite number of adjacent states, encompassing the whole fixed-lag period. Hence, the expectation step (E-step) of our algorithm comprises a diagonalized DFT-domain Kalman filter [6] followed by a DFT-domain Kalman smoother, which also maintains diagonality. The smoother stage becomes imperative for learning the complete posterior and achieving the desired convergence [2]. The maximization step (M-step) then exploits the state estimates and pertinent expectations evaluated in the E-step to yield optimum model parameters in the maximum-likelihood sense for each block-time index in the lag interval.

In Sec. 2 we describe the DFT-domain observation and state-transition equations, which together represent the dynamical model. The focus of Sec. 3 is to express the complete data likelihood for a fixed-lag interval and derive the corresponding ML-EM algorithm. In Sec. 4 we delve into simulation results to ascertain the performance of the algorithm under diverse conditions of noise and state-variability. Conclusions are presented in Sec. 5.

We use nonbold lowercase letters for scalar quantities, bold lowercase for vectors, and bold uppercase for frequency-domain quantities. The superscript H denotes Hermitian transposition. We use $\mathbf{F}_M$ and log to denote the DFT matrix of size M and the natural logarithm, respectively. Lowercase letters "t" and "τ" are reserved for sample- and block-time indices, while R and N are the block-shift in samples and the number of blocks in the fixed-lag interval, respectively.

2. DFT-DOMAIN DYNAMICAL MODELING

We address the system identification problem, where the time-domain input x_t gets convolved through the unknown FIR system $\mathbf{w}_\tau$ to give an intermediate signal d_t. Measurement noise s_t is superimposed on the intermediate signal to give the observation y_t. The estimated state of the unknown system is denoted by $\widehat{\mathbf{w}}_\tau$.

2.1. Markov Model of a Time-Varying System

A DFT-domain state vector $\mathbf{W}_\tau$ is defined on the basis of the constraint that a time-domain state $\mathbf{w}_\tau$ of $M-R$ non-zero coefficients is represented, i.e., $\mathbf{W}_\tau = \mathbf{F}_M \begin{bmatrix}\mathbf{w}_\tau^H & \mathbf{0}\end{bmatrix}^H$, where M is the block-size. Thus, the first-order Markov model expressing the evolution of the state vector $\mathbf{W}_\tau$ is described as:

$$\mathbf{W}_\tau = A \cdot \mathbf{W}_{\tau-1} + \Delta\mathbf{W}_\tau \,. \tag{1}$$

Here, the factor A denotes a state transition coefficient in the range $0 < A < 1$ and $\Delta\mathbf{W}_\tau$ represents zero-mean and block-wise un-

 ISSN 2076-1465

correlated process noise with $M \times M$ process noise covariance matrix $\mathrm{E}\{\Delta\mathbf{W}_\tau \Delta\mathbf{W}_\tau^H\} = \Psi_\tau^\Delta$.

2.2. Block-Frequency-Domain Observation Model

We define block-oriented input and measurement noise signals as

$$\begin{aligned}\mathbf{x}_\tau &= [x_{\tau R-M+1}, x_{\tau R-M+2}, ..., x_{\tau R}]^H \\ \mathbf{s}_\tau &= [s_{\tau R-R+1}, s_{\tau R-R+2}, ..., s_{\tau R}]^H \ .\end{aligned} \tag{2}$$

An $M \times M$ input matrix $\mathbf{X}_\tau$ is created by first applying DFT and then diagonalization to the input signal, i.e., $\mathbf{X}_\tau = \mathrm{diag}\{\mathbf{F}_M \mathbf{x}_\tau\}$. Using these definitions, we apply an overlap-save convolution to obtain an observation vector $\mathbf{y}_\tau$ defined analogous to $\mathbf{s}_\tau$,

$$\mathbf{y}_\tau = \mathbf{Q}^H \mathbf{F}_M^{-1} \mathbf{X}_\tau \mathbf{W}_\tau + \mathbf{s}_\tau \ , \tag{3}$$

where $\mathbf{Q}^H = (\mathbf{0} \quad \mathbf{I}_R)$ is an $R \times M$ projection matrix, included to linearize the cyclic convolution in DFT-domain, and $\mathbf{I}_R$ denotes an $R \times R$ identity. Application of $\mathbf{Q}$ and $\mathbf{F}_M$ to (3) results in a DFT-domain representation $\mathbf{Y}_\tau = \mathbf{F}_M \mathbf{Q} \mathbf{y}_\tau$ of the observation vector:

$$\mathbf{Y}_\tau = \mathbf{F}_M \mathbf{Q} \mathbf{Q}^H \mathbf{F}_M^{-1} \mathbf{X}_\tau \mathbf{W}_\tau + \mathbf{F}_M \mathbf{Q} \mathbf{s}_\tau \ . \tag{4}$$

The constant term $\mathbf{G} = \mathbf{F}_M \mathbf{Q} \mathbf{Q}^H \mathbf{F}_M^{-1}$ can be combined with the input matrix $\mathbf{X}_\tau$ to give the overlap-save constrained version $\mathbf{C}_\tau = \mathbf{G}\mathbf{X}_\tau$ of the input signal. These abbreviations then allow a more compact matrix-vector respresentation of the observation model,

$$\mathbf{Y}_\tau = \mathbf{C}_\tau \mathbf{W}_\tau + \mathbf{S}_\tau \ , \tag{5}$$

where $\mathbf{S}_\tau = \mathbf{F}_M \mathbf{Q} \mathbf{s}_\tau$ is considered as zero-mean and block-wise uncorrelated measurement noise with $M \times M$ time-varying measurement noise covariance $\mathrm{E}\{\mathbf{S}_\tau \mathbf{S}_\tau^H\} = \Psi_\tau^S$.

Combining the Markov model in (1) with the observation equation in (5), we obtain a DFT-domain state-space model. The covariances $\Theta_\tau = \{\Psi_\tau^S, \Psi_\tau^\Delta\}$ are then termed as the unknown parameters of the state-space model.

2.3. Dynamical Model Over the Fixed-Lag Interval

From the equations of the dynamical model, (1) and (5), we express the i-th DFT-domain block in a given lag interval indexed with κ,

$$\begin{aligned}\mathbf{W}_{\kappa,i} &= A \cdot \mathbf{W}_{\kappa,i-1} + \Delta\mathbf{W}_{\kappa,i} \\ \mathbf{Y}_{\kappa,i} &= \mathbf{C}_{\kappa,i} \mathbf{W}_{\kappa,i} + \mathbf{S}_{\kappa,i} \ .\end{aligned} \tag{6}$$

where $\mathbf{W}_{\kappa,i} = \mathbf{W}_\tau$, analogy of which can be extended to all other quantities in (6). The block-index τ and the lag-interval number κ are related via $\tau = \kappa N + i$ and $i = 1, 2, \cdots, N$. A lag interval corresponds to a non-overlapping window comprising N DFT-domain blocks. The corresponding parameters take the form $\Theta_{\kappa,i} = \left\{\Psi_{\kappa,i}^S, \Psi_{\kappa,i}^\Delta\right\}$. Expressions in (6) are illustrated in Fig. 1, which highlights the state-transition and the observation strata, and depicts the scope of a fixed-lag.

3. ML-EM ALGORITHM

Maximum-likelihood learning of distinct model parameters $\Theta_{\kappa,1:N}$ for each block-index τ entails the maximization of the following complete data log-likelihood function over a given fixed-lag interval:

$$\begin{aligned}\mathscr{L}(\Theta_{\kappa,1:N}) &= \log p(\mathbf{Y}_{\kappa,1:N} | \Theta_{\kappa,1:N}) \\ &= \log \int p(\mathbf{Y}_{\kappa,1:N}, \mathbf{W}_{\kappa,1:N} | \Theta_{\kappa,1:N}) \, d\mathbf{W}_{\kappa,1:N} \ .\end{aligned} \tag{7}$$

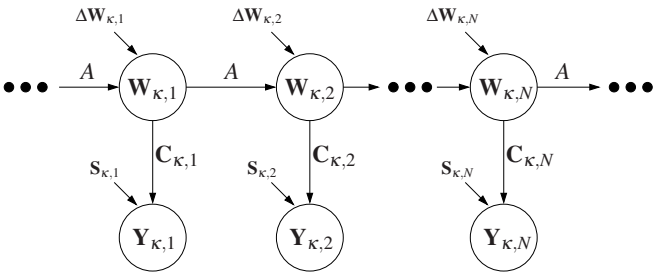

Fig. 1. DFT-domain dynamical modeling over a fixed-lag interval.

The term $p(\mathbf{Y}_{\kappa,1:N}, \mathbf{W}_{\kappa,1:N} | \Theta_{\kappa,1:N})$ denotes a joint distribution conditioned on model parameters $\Theta_{\kappa,1:N}$, and on the implicitly assumed known initial state predictor $\mathbf{W}_{\kappa,0}$. Considering an arbitrary distribution $q(\mathbf{W}_{\kappa,1:N})$, irrespective of its form, we can obtain a lower bound $\mathscr{F}(q, \Theta_{\kappa,1:N})$ on the log-likelihood,

$$\begin{aligned}\mathscr{L}(\Theta_{\kappa,1:N}) &= \log \int q(\mathbf{W}_{\kappa,1:N}) \, \frac{p(\mathbf{Y}_{\kappa,1:N}, \mathbf{W}_{\kappa,1:N} | \Theta_{\kappa,1:N})}{q(\mathbf{W}_{\kappa,1:N})} \, d\mathbf{W}_{\kappa,1:N} \\ &\geq \int q(\mathbf{W}_{\kappa,1:N}) \log \frac{p(\mathbf{Y}_{\kappa,1:N}, \mathbf{W}_{\kappa,1:N} | \Theta_{\kappa,1:N})}{q(\mathbf{W}_{\kappa,1:N})} \, d\mathbf{W}_{\kappa,1:N} \\ &= \mathscr{F}(q, \Theta_{\kappa,1:N}),\end{aligned} \tag{8}$$

making use of Jensen's inequality, signifying the concavity of the log function. The EM algorithm alternates between maximizing the functional $\mathscr{F}(q, \Theta_{\kappa,1:N})$ with respect to the distribution $q(\mathbf{W}_{\kappa,1:N})$ and the model parameters $\Theta_{\kappa,1:N}$ [8]. EM iterations can be expressed as:

$$\begin{aligned}\textbf{E-step}: \quad & q^k \longleftarrow \arg\max_q \mathscr{F}(q, \Theta_{\kappa,1:N}^{k-1}) \\ \textbf{M-step}: \quad & \Theta_{\kappa,1:N}^k \longleftarrow \arg\max_\Theta \mathscr{F}(q^k, \Theta_{\kappa,1:N}) \ ,\end{aligned} \tag{9}$$

where k denotes the iteration index.

3.1. Expectation Step

The joint distribution $p(\mathbf{Y}_{\kappa,1:N}, \mathbf{W}_{\kappa,1:N} | \Theta_{\kappa,1:N})$ used in (8) can be factorized using the Bayes' theorem into the posterior $p(\mathbf{W}_{\kappa,1:N} | \mathbf{Y}_{\kappa,1:N}, \Theta_{\kappa,1:N})$ and the likelihood $p(\mathbf{Y}_{\kappa,1:N} | \Theta_{\kappa,1:N})$:

$$\begin{aligned}p(\mathbf{Y}_{\kappa,1:N}, \mathbf{W}_{\kappa,1:N} | \Theta_{\kappa,1:N}) = \\ p(\mathbf{W}_{\kappa,1:N} | \mathbf{Y}_{\kappa,1:N}, \Theta_{\kappa,1:N}) \, p(\mathbf{Y}_{\kappa,1:N} | \Theta_{\kappa,1:N}).\end{aligned} \tag{10}$$

On substituting (10) into (8), we can see that maximization in the E-step is achieved by setting $q(\mathbf{W}_{\kappa,1:N}) = p(\mathbf{W}_{\kappa,1:N} | \mathbf{Y}_{\kappa,1:N}, \Theta_{\kappa,1:N})$, because for such a selection of $q(\mathbf{W}_{\kappa,1:N})$ the objective functional $\mathscr{F}(q, \Theta_{\kappa,1:N})$ equals the intermediate log-likelihood at iteration time k [1]. The sample posterior for a given lag interval $p(\mathbf{W}_{\kappa,i} | \mathbf{Y}_{\kappa,1:N}, \Theta_{\kappa,1:N})$ can be recursively learned by means of a Kalman filter-smoother combination [9].

3.1.1. DFT-Domain Adaptive Kalman Filter

Owing to the assumption of Gaussianity, the mean and covariance of the filtering distribution $p(\mathbf{W}_{\kappa,i} | \mathbf{Y}_{\kappa,i}, \Theta_{\kappa,1:N})$ in the E-step are computed by recursive equations of the DFT-domain Kalman filter which have been formulated in [6]. The Kalman filtering stage, which learns the filtering distribution $p(\mathbf{W}_{\kappa,i} | \mathbf{Y}_{\kappa,i}, \Theta_{\kappa,1:N})$ constitutes a *forward pass* in terms of the processing direction of the algorithm, as shown in Fig. 2. It has been verified in [6, 10] that $\mathbf{G}$, cf. Sec. 2.3, can be approximated as a scaled identity, $\mathbf{G} \approx \frac{R}{M} \mathbf{I}_M$, and thus $\mathbf{C}_{\kappa,i} \approx \frac{R}{M} \mathbf{X}_{\kappa,i}$. Considering this and the approximate diagonality of the covariances $\Psi_{\kappa,i}^S$ and $\Psi_{\kappa,i}^\Delta$ in DFT-domain,

the exact Kalman filter recursions for a given lag interval can be approximated as a diagonalized DFT-domain adaptive Kalman filter [6]:

for $i = 1$ to N
  if$(i == 1)$
    $\widehat{\mathbf{W}}_{\kappa,i}^{\,i-1} \leftarrow \mathbf{W}_{\kappa,0}$
    $\mathbf{P}_{\kappa,i}^{\,i-1} \leftarrow \mathbf{P}_{\kappa,0}$
  else
    $\widehat{\mathbf{W}}_{\kappa,i}^{\,i-1} \leftarrow A \cdot \widehat{\mathbf{W}}_{\kappa,i-1}^{\,i-1}$
    $\mathbf{P}_{\kappa,i}^{\,i-1} \leftarrow A^2 \cdot \mathbf{P}_{\kappa,i-1}^{\,i-1} + \Psi_{\kappa,i}^{\Delta}$
  end if
  $\mu_{\kappa,i} \leftarrow \mathbf{P}_{\kappa,i}^{\,i-1} \left[\mathbf{X}_{\kappa,i} \mathbf{P}_{\kappa,i}^{\,i-1} \mathbf{X}_{\kappa,i}^{H} + \frac{M}{R} \Psi_{\kappa,i}^{S} \right]^{-1}$
  $\mathbf{E}_{\kappa,i} \leftarrow \mathbf{Y}_{\kappa,i} - \mathbf{C}_{\kappa,i} \widehat{\mathbf{W}}_{\kappa,i}^{\,i-1}$
  $\widehat{\mathbf{W}}_{\kappa,i}^{\,i} \leftarrow \widehat{\mathbf{W}}_{\kappa,i}^{\,i-1} + \mu_{\kappa,i} \mathbf{X}_{\kappa,i}^{H} \mathbf{E}_{\kappa,i}$
  $\mathbf{P}_{\kappa,i}^{\,i} \leftarrow \left[\mathbf{I}_M - \frac{R}{M} \mu_{\kappa,i} \mathbf{X}_{\kappa,i}^{H} \mathbf{X}_{\kappa,i} \right] \mathbf{P}_{\kappa,i}^{\,i-1}$
end for

Here, $\mu_{\kappa,i}$ is the Kalman step-size, $\mathbf{E}_{\kappa,i}$ the error signal, $\widehat{\mathbf{W}}_{\kappa,i}^{\,i}$ the estimate of the unknown state, $\mathbf{P}_{\kappa,i}^{\,i}$ the state error covariance, and $\mathbf{I}_M$ the identity. The intermediate terms, $\widehat{\mathbf{W}}_{\kappa,i}^{\,i-1}$ and $\widehat{\mathbf{P}}_{\kappa,i}^{\,i-1}$, represent one-step predictors for the estimate of the unknown state and the state error covariance. Our framework assumes the knowledge of the initial belief state, hence the initial predictors for the state $\mathbf{W}_{\kappa,0}$ and error covariance $\mathbf{P}_{\kappa,0}$ are considered known from the previous lag interval.

3.1.2. DFT-Domain Fixed-Interval Smoother

We augment the DFT-domain Kalman filter with the DFT-domain smoother, which constitutes the *backward pass*, to accomplish the learning of the complete data posterior $p(\mathbf{W}_{\kappa,i}|\mathbf{Y}_{\kappa,1:N}, \Theta_{\kappa,1:N})$ from the already learned filtering distribution $p(\mathbf{W}_{\kappa,i}|\mathbf{Y}_{\kappa,i}, \Theta_{\kappa,1:N})$ [9]:

for $i = N$ down to 2
  $\mathbf{J}_{\kappa,i-1} \leftarrow \mathbf{P}_{\kappa,i-1}^{\,i-1} \cdot A \cdot \left[\mathbf{P}_{\kappa,i}^{\,i-1} \right]^{-1}$
  $\widehat{\mathbf{W}}_{\kappa,i-1}^{\,N} \leftarrow \widehat{\mathbf{W}}_{\kappa,i-1}^{\,i-1} + \mathbf{J}_{\kappa,i-1} \left[\widehat{\mathbf{W}}_{\kappa,i}^{\,N} - \widehat{\mathbf{W}}_{\kappa,i}^{\,i-1} \right]$
  $\mathbf{P}_{\kappa,i-1}^{\,N} \leftarrow \mathbf{P}_{\kappa,i-1}^{\,i-1} + \mathbf{J}_{\kappa,i-1} \left[\mathbf{P}_{\kappa,i}^{\,N} - \mathbf{P}_{\kappa,i}^{\,i-1} \right] \mathbf{J}_{\kappa,i-1}^{H}$
end for

The term $\mathbf{J}_{\kappa,i-1}$ is the smoother gain, while $\widehat{\mathbf{W}}_{\kappa,i}^{\,N}$ and $\mathbf{P}_{\kappa,i}^{\,N}$ are the mean and covariance of the smoothed posterior. It follows from the direct reliance of the smoother gain $\mathbf{J}_{\kappa,i-1}$ on the state error covariance $\mathbf{P}_{\kappa,i-1}^{\,i-1}$ and its predictor $\mathbf{P}_{\kappa,i}^{\,i-1}$ that the smoother quantities inherit the diagonal attributes of the filter in DFT-domain. Fig. 2 while depicting both the forward and the backward passes, conceptually illustrates the estimation of $\widehat{\mathbf{W}}_{\kappa,i}^{\,N}$ and the subsequent and indispensable interaction with the M-step for parameter learning.

3.2. M-Step: Parameter Learning Rules

The lower bound on the log-likelihood contains a term which is independent of the model parameters $\Theta_{\kappa,1:N}$:

$$\mathscr{F}(q, \Theta_{\kappa,1:N}) = \int q(\mathbf{W}_{\kappa,1:N}) \log p(\mathbf{Y}_{\kappa,1:N}, \mathbf{W}_{\kappa,1:N} | \Theta_{\kappa,1:N}) \, d\mathbf{W}_{\kappa,1:N} - \int q(\mathbf{W}_{\kappa,1:N}) \log q(\mathbf{W}_{\kappa,1:N}) \, d\mathbf{W}_{\kappa,1:N} . \quad (11)$$

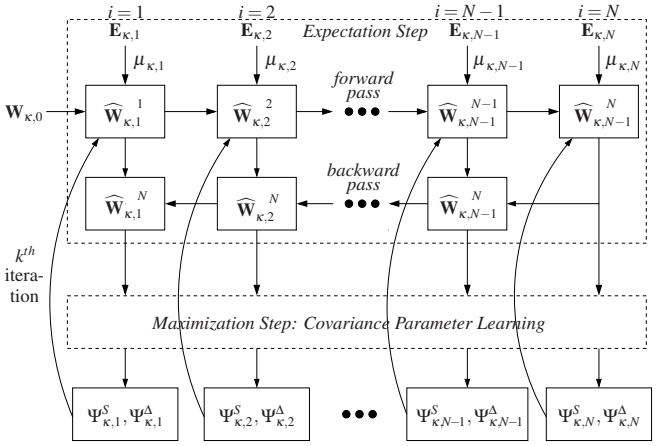

Fig. 2. DFT-domain expectation maximization algorithm for block-wise parameter learning.

Since the second term, the *entropy* of $q(\mathbf{W}_{\kappa,1:N})$, is not a function of $\Theta_{\kappa,1:N}$, the k-th M-step is computed by maximizing the first term only:

$$\Theta_{\kappa,1:N}^{k} = \arg\max_{\Theta_{\kappa,1:N}} \int q^{k}(\mathbf{W}_{\kappa,1:N}) \log p(\mathbf{Y}_{\kappa,1:N}, \mathbf{W}_{\kappa,1:N} | \Theta_{\kappa,1:N}) \, d\mathbf{W}_{\kappa,1:N} . \quad (12)$$

The joint distribution in (12) can be factorized to highlight observation and state-transition distributions:

$$p(\mathbf{Y}_{\kappa,1:N}, \mathbf{W}_{\kappa,1:N} | \Theta_{\kappa,1:N}) = \prod_{i=1}^{N} p(\mathbf{Y}_{\kappa,i} | \mathbf{W}_{\kappa,i}, \Theta_{\kappa,i}) \; p(\mathbf{W}_{\kappa,i} | \mathbf{W}_{\kappa,i-1}, \Theta_{\kappa,i}) . \quad (13)$$

As the state-space model has been formulated in DFT-domain, both transition and transmission processes are described by complex multivariate Gaussians [11], i.e., we express

$$p(\mathbf{Y}_{\kappa,1:N}, \mathbf{W}_{\kappa,1:N} | \Theta_{\kappa,1:N}) = \prod_{i=1}^{N} \frac{1}{\pi^{M} |\Psi_{\kappa,i}^{S}|} \exp\left[-\left(\mathbf{Y}_{\kappa,i} - \mathbf{C}_{\kappa,i} \mathbf{W}_{\kappa,i}\right)^{H} \Psi_{\kappa,i}^{S^{-1}} \left(\mathbf{Y}_{\kappa,i} - \mathbf{C}_{\kappa,i} \mathbf{W}_{\kappa,i}\right) \right] \times \prod_{i=1}^{N} \frac{1}{\pi^{M} |\Psi_{\kappa,i}^{\Delta}|} \exp\left[-\left(\mathbf{W}_{\kappa,i} - A \cdot \mathbf{W}_{\kappa,i-1}\right)^{H} \Psi_{\kappa,i}^{\Delta^{-1}} \left(\mathbf{W}_{\kappa,i} - A \cdot \mathbf{W}_{\kappa,i-1}\right) \right] , \quad (14)$$

where $|\cdot|$ denotes the determinant of an $M \times M$ matrix. In accordance with (12), we apply the log function to (14) followed by an expectation operation with respect to the distribution $q^{k}(\mathbf{W}_{\kappa,1:N})$, i.e.,

$$\mathrm{E}_{q^{k}}\{\log p(\mathbf{Y}_{\kappa,1:N}, \mathbf{W}_{\kappa,1:N} | \Theta_{\kappa,1:N})\} = \sum_{i=1}^{N} \Big\{ -2M \log\pi - \log|\Psi_{\kappa,i}^{S}| - \log|\Psi_{\kappa,i}^{\Delta}| - \mathrm{Tr}\Big\{ \Psi_{\kappa,i}^{S^{-1}} \mathrm{E}_{q^{k}}\Big\{ \left(\mathbf{Y}_{\kappa,i} - \mathbf{C}_{\kappa,i} \mathbf{W}_{\kappa,i}\right) \left(\mathbf{Y}_{\kappa,i} - \mathbf{C}_{\kappa,i} \mathbf{W}_{\kappa,i}\right)^{H} \Big\} \Big\} - \mathrm{Tr}\Big\{ \Psi_{\kappa,i}^{\Delta^{-1}} \mathrm{E}_{q^{k}}\Big\{ \left(\mathbf{W}_{\kappa,i} - A \cdot \mathbf{W}_{\kappa,i-1}\right) \left(\mathbf{W}_{\kappa,i} - A \cdot \mathbf{W}_{\kappa,i-1}\right)^{H} \Big\} \Big\} \Big\} , \quad (15)$$

where $\mathrm{Tr}\{\cdot\}$ is the trace operator. Covariance terms $\Psi_{\kappa,i}^{S}$ and $\Psi_{\kappa,i}^{\Delta}$, for $i = 1$ to N, are estimated by taking the corresponding partial

derivatives of these expressions and setting them to zero. Using the lemmata [12]

$$\frac{\partial \log|\mathbf{U}|}{\partial \mathbf{U}} = \mathbf{U}^{-T},\ \frac{\partial \mathbf{U}^{-1}}{\partial \mathbf{U}} = -\mathbf{U}^{-1}\mathbf{U}^{-1},\ \frac{\partial \mathrm{Tr}\{\mathbf{U}\,\mathbf{V}\}}{\partial \mathbf{U}} = \mathbf{V}^{T}\ , \quad (16)$$

we find the following expressions for i-th parameters in the k-th EM iteration that maximize (15):

$$\Psi^{S\,k}_{\kappa,i} = \mathrm{E}_{q^k}\Big\{ \big(\mathbf{Y}_{\kappa,i} - \mathbf{C}_{\kappa,i}\mathbf{W}_{\kappa,i}\big)\big(\mathbf{Y}_{\kappa,i} - \mathbf{C}_{\kappa,i}\mathbf{W}_{\kappa,i}\big)^{H} \Big\}\ , \quad (17)$$

$$\Psi^{\Delta\,k}_{\kappa,i} = \mathrm{E}_{q^k}\Big\{ \big(\mathbf{W}_{\kappa,i} - A\cdot\mathbf{W}_{\kappa,i-1}\big)\big(\mathbf{W}_{\kappa,i} - A\cdot\mathbf{W}_{\kappa,i-1}\big)^{H} \Big\}\ . \quad (18)$$

The result in (18) represents learning rules that are optimal in the ML sense under the inferred distribution $q_k(\mathbf{W}_{\kappa,1:N})$.

3.2.1. Evaluation of the Measurement Noise Covariance

The learning rule for the measurement noise covariance $\Psi^{S\,k}_{\kappa,i}$ in (17) can be further expanded as:

$$\begin{aligned}\Psi^{S\,k}_{\kappa,i} = {} & \mathbf{Y}_{\kappa,i}\mathbf{Y}^{H}_{\kappa,i} - \mathbf{Y}_{\kappa,i}\mathrm{E}_{q^k}\{\mathbf{W}^{N\,H}_{\kappa,i}\}\mathbf{C}^{H}_{\kappa,i} \\ & - \mathbf{C}_{\kappa,i}\mathrm{E}_{q^k}\Big\{\mathbf{W}_{\kappa,i}{}^{N}\Big\}\mathbf{Y}^{H}_{\kappa,i} + \mathbf{C}_{\kappa,i}\mathrm{E}_{q^k}\{\mathbf{W}_{\kappa,i}{}^{N}\mathbf{W}^{N\,H}_{\kappa,i}\}\mathbf{C}^{H}_{\kappa,i}\ .\end{aligned} \quad (19)$$

It is evident from (19) that the expectations $\mathrm{E}_{q^k}\Big\{\mathbf{W}_{\kappa,i}\mathbf{W}^{H}_{\kappa,i}\Big\}$ and $\mathrm{E}_{q^k}\big\{\mathbf{W}_{\kappa,i}\big\}$ have to be evaluated. The mean $\widehat{\mathbf{W}}_{\kappa,i}{}^{N}$ and covariance $\mathbf{P}_{\kappa,i}{}^{N}$ obtained from the DFT-domain adaptive Kalman filter-smoother (E-step) allow us to write

$$\begin{aligned}\mathrm{E}_{q^k}\big\{\mathbf{W}_{\kappa,i}\big\} &= \widehat{\mathbf{W}}_{\kappa,i}{}^{N} \\ \mathrm{E}_{q^k}\Big\{\mathbf{W}_{\kappa,i}\mathbf{W}^{H}_{\kappa,i}\Big\} &= \widehat{\mathbf{W}}_{\kappa,i}{}^{N}\widehat{\mathbf{W}}^{N\,H}_{\kappa,i} + \mathbf{P}_{\kappa,i}{}^{N}\ .\end{aligned} \quad (20)$$

Using the expectation in (20) we simplify the expression in (19) as,

$$\Psi^{S}_{\kappa,i} = \mathbf{E}^{+}_{\kappa,i}\mathbf{E}^{+H}_{\kappa,i} + \mathbf{C}_{\kappa,i}\mathbf{P}_{\kappa,i}{}^{N}\mathbf{C}^{H}_{\kappa,i}\ , \quad (21)$$

where $\mathbf{E}^{+}_{\kappa,i} = \mathbf{Y}_{\kappa,i} - \mathbf{C}_{\kappa,i}\widehat{\mathbf{W}}_{\kappa,i}{}^{N}$ is the *a posteriori* error, while $\mathbf{C}_{\kappa,i}\mathbf{P}_{\kappa,i}{}^{N}\mathbf{C}^{H}_{\kappa,i}$ can be approximated with $\frac{R}{M}\mathbf{X}_{\kappa,i}\mathbf{P}_{\kappa,i}{}^{N}\mathbf{X}^{H}_{\kappa,i}$ [6].

3.2.2. Evaluation of the Process Noise Covariance

The expansion of the quantity $\Psi^{\Delta\,k}_{\kappa,i}$ in (18) leads to

$$\begin{aligned}\Psi^{\Delta\,k}_{\kappa,i} = {} & \mathrm{E}_{q^k}\Big\{\mathbf{W}_{\kappa,i}\mathbf{W}^{H}_{\kappa,i}\Big\} - A\cdot\mathrm{E}_{q^k}\Big\{\mathbf{W}_{\kappa,i}\mathbf{W}^{H}_{\kappa,i-1}\Big\} \\ & - A\cdot\mathrm{E}_{q^k}\Big\{\mathbf{W}_{\kappa,i-1}\mathbf{W}^{H}_{\kappa,i}\Big\} + A^2\cdot\mathrm{E}_{q^k}\Big\{\mathbf{W}_{\kappa,i-1}\mathbf{W}^{H}_{\kappa,i-1}\Big\}\ ,\end{aligned} \quad (22)$$

which requires the evaluation of the expectation $\mathrm{E}_{q^k}\Big\{\mathbf{W}_{\kappa,i}\mathbf{W}^{H}_{\kappa,i-1}\Big\}$. This term can be expressed according to [9] as:

$$\mathrm{E}_{q^k}\Big\{\mathbf{W}_{\kappa,i}\mathbf{W}^{H}_{\kappa,i-1}\Big\} = \widehat{\mathbf{W}}_{\kappa,i}{}^{N}\widehat{\mathbf{W}}^{N\,H}_{\kappa,i-1} + \mathbf{P}_{\kappa,i,i-1}{}^{N}\ . \quad (23)$$

To obtain the above expectation we therefore invoke the *lag-one covariance smoother* [9], which indirectly relies on the computations of the E-step state-estimator to give, for $i = N-1, N-2, \cdots, 2$:

$$\mathbf{P}_{\kappa,i,i-1}{}^{N} = \mathbf{P}_{\kappa,i}{}^{i}\mathbf{J}_{\kappa,i-1} + \mathbf{J}_{\kappa,i}\big(\mathbf{P}_{\kappa,i+1,i}{}^{N} - A\cdot\mathbf{P}_{\kappa,i}{}^{i}\big)\mathbf{J}_{\kappa,i-1}\ . \quad (24)$$

To bring lag-one covariance smoother in conformity with the aforementioned diagonalization, we initialize for $i = N$ as:

$$\mathbf{P}_{\kappa,N,N-1}{}^{N} = \left[\mathbf{I}_M - \frac{R}{M}\mu_{\kappa,N}\mathbf{X}^{H}_{\kappa,N}\mathbf{X}_{\kappa,N}\right]\mathbf{P}_{\kappa,N-1}{}^{N-1}\cdot A\ . \quad (25)$$

In light of the expectations, as evaluated in (20) and (23), we can proceed from (22) to obtain the learning rule for the process noise covariance,

$$\Psi^{\Delta\,k}_{\kappa,i} = \Delta\widehat{\mathbf{W}}_{\kappa,i}{}^{N}\Delta\widehat{\mathbf{W}}^{N\,H}_{\kappa,i} + \mathbf{P}_{\kappa,i}{}^{N} + A^2\cdot\mathbf{P}_{\kappa,i-1}{}^{N} - 2\cdot A\cdot\mathbf{P}_{\kappa,i,i-1}{}^{N}\ , \quad (26)$$

where $\Delta\widehat{\mathbf{W}}_{\kappa,i}{}^{N} = \widehat{\mathbf{W}}_{\kappa,i}{}^{N} - A\cdot\widehat{\mathbf{W}}_{\kappa,i-1}{}^{N}$. It is worth mentioning that owing to the assumption of diagonality on $\Psi^{S}_{\kappa,i}$ and $\Psi^{\Delta}_{\kappa,i}$, only the main diagonals of $\mathbf{E}^{+}_{\kappa,i}\mathbf{E}^{+H}_{\kappa,i}$ and $\Delta\widehat{\mathbf{W}}_{\kappa,i}{}^{N}\Delta\widehat{\mathbf{W}}^{N\,H}_{\kappa,i}$ are evaluated in (21) and (26), respectively.

4. RESULTS

For analyzing the performance of the derived algorithm we have considered dynamical systems with quantifiable variability under diverse conditions of observation noise. The dynamical systems used for simulations conform to Markov model characteristics, cf. Sec. 2.1, with a transition coefficient A and a corresponding time-constant $\rho = -R/(f_s\cdot\log A)$, where $f_s = 16$ kHz is the sampling frequency. Block-size M and block-shift R of the DFT-domain model were set to 256 and 64, respectively. The observation noise signal, which acquires the form of a burst noise signal, provides for time-varying SNR conditions so that the online covariance estimation capabilities of the algorithm can be ascertained. The dynamic range of the observation noise covariance is 25 dB.

We initiated our analysis by evaluating the learning of the time-varying observation noise covariance Ψ^{S}_{τ}. The data for this evaluation was derived from a Markov model with a large time-constant of $\rho = 40$s, i.e., $A = 0.9999$. Fig. 3 illustrates the obtained values of Ψ^{S}_{τ} along with the true observation noise covariance Ψ^{S}_{τ}. The proposed algorithm was compared with the conventional approach of performing averaging of parameters in the M-step. As the averaging in M-step is performed over whole of the fixed-lag duration, the conjecture of stationarity of the underlying model is inevitable for the given interval. We have considered M-step averaging scheme [13] with various values of the fixed-lag interval, ranging from $N = 100$ to $N = 10$ blocks. It can be observed that the inferred Ψ^{S}_{τ} approaches closer to the true covariance Ψ^{S}_{τ} as the fixed-lag interval is reduced from $N = 100$ to $N = 10$, which can be expected as the length of *event* in the form of noise burst is about 50 blocks. The proposed approach with $N = 100$ is, in contrast, oblivious to the length of the fixed-lag interval and provides near perfect tracking of the observation noise covariance. Fig. 4 illustrates the effect of the quality of observation noise covariance learning on system identification. Relative system distance D was considered as the instrumental measure for quantifying system identification, i.e., $D = 10\log_{10}(\mathrm{E}\{||\mathbf{W}_{\tau} - \widehat{\mathbf{W}}_{\tau}||^2\}/\mathrm{E}\{||\mathbf{W}_{\tau}||^2\})$. It can be observed in conjunction with Fig. 3 that the proposed algorithm on the average, where instantaneous estimates are provided to the E-step, achieves the lowest system distance D. Noticeable in the figure is also the degradation in system identification caused by averaging in the M-step in the contending configurations, which originates from inadequate parameter learning.

We carried out a similar analysis for evaluating the learning of the process noise covariance Ψ^{Δ}_{τ}. In order to focus the analysis on the learning of the process noise covariance, evaluations were carried out under low and constant observation noise conditions. The data for the analysis was derived by switching between two Markov models, with time constants of $\rho = 4$s and $\rho = 45$s corresponding to

$A = 0.999$ and $A = 0.9999$, respectively. As the simulation data has been generated using an underlying system with unit state covariance $\Psi_\tau^W = E\{\mathbf{W}_\tau \mathbf{W}_\tau^H\}$, the true value of Ψ_τ^Δ can be derived under the stationarity constraint from (1) as:

$$\begin{aligned}\Psi_\tau^\Delta &= (1-A^2)\cdot\Psi_\tau^W\ , \\ &= (1-A^2)\cdot 1\ . \end{aligned} \tag{27}$$

Hence, the true value of Ψ_τ^Δ, as depicted in Fig. 5, switches between -27 dB and -37 dB for $A = 0.999$ and $A = 0.9999$, respectively. A larger value of Ψ_τ^Δ signifies an underlying model with a smaller time-constant and vice versa. It can be seen in the plot that all competitors seek to track the process noise covariance of the underlying model, but the smoothing approach with $N = 100$ almost completely fails to detect the switching process. Cases with $N = 50$ and $N = 10$ tend to track better, but suffer from over and under-estimation, respectively. The learning curve for the proposed algorithm, despite larger variance, effectively recognizes the switching event and veritably tracks the model covariance Ψ_τ^Δ.

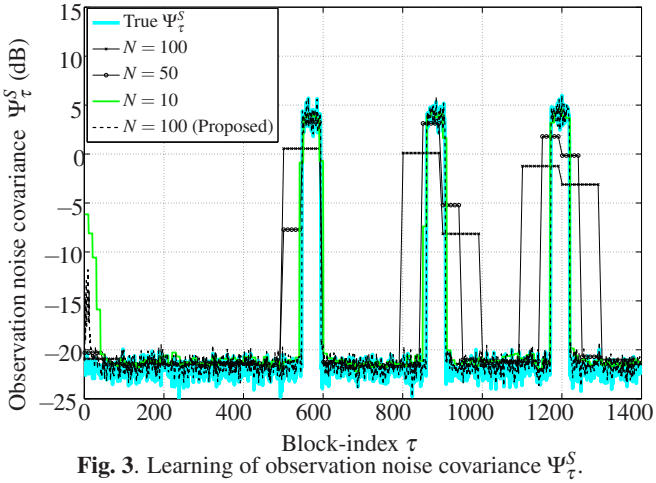

Fig. 3. Learning of observation noise covariance Ψ_τ^S.

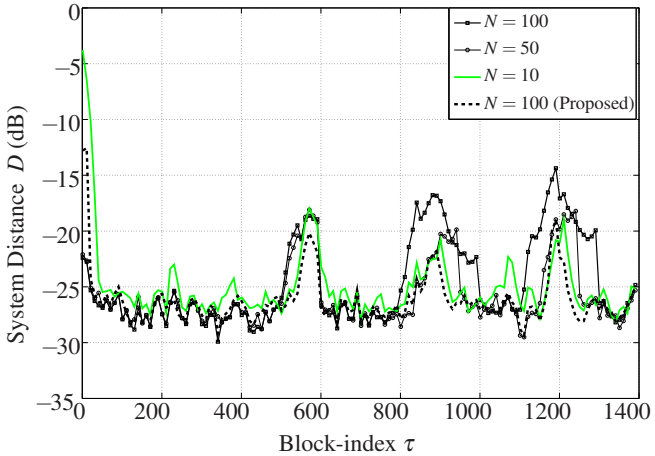

Fig. 4. Impact of parameter smoothing on system identification.

5. CONCLUSIONS

We have derived a DFT-domain maximum-likelihood expectation-maximization algorithm to carry out joint state and parameter learning of linear time-varying dynamical models. Our framework puts forth an efficient diagonalized DFT-domain Kalman filter-smoother

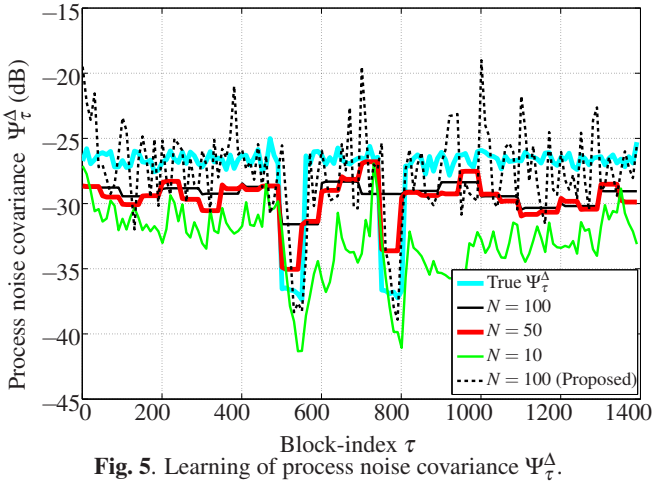

Fig. 5. Learning of process noise covariance Ψ_τ^Δ.

state estimator, which manifests the expectation step iteratively with a maximization step. The maximization step utilizes the quantities estimated in the expectation step to output distinct model parameters for each DFT-domain block within the fixed-lag interval. We have thus shown that the fixed-lag smoothing in the expectation step does not necessarily entail averaging in the maximization step. This enables the algorithm to effectively adapt even in non-stationary environments. We have substantiated the aforementioned feature of our algorithm by means of simulation results and comparison to a maximization step averaging approach.

6. REFERENCES

[1] M. Beal, *Variational Algorithms for Approximate Bayesian Inference*, PhD. Thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[2] K. Murphy, *Dynamic Bayesian Networks: Representation, Inference and Learning*, Dissertation, University of California, Berkeley, 2002.

[3] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, NJ, 2002.

[4] S. Haykin, A. Sayed, J. Zeidler, P. Yee, and P. Wei, "Adaptive tracking of linear time-variant systems by extended RLS algorithms," *IEEE Trans. on Signal Processing*, vol. 45, pp. 1118–1127, May 1997.

[5] S. Roweis and Z. Ghahramani, "A unifying review of linear Gaussian models," *Neural Computation*, 11(2), 1999.

[6] G. Enzner and P. Vary, "Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones," *Signal Processing, Elsevier*, vol. 86(6), pp. 1140–1156, Jun. 2006.

[7] R. Neal and G. Hinton, "A view of EM algorithm that justifies incremental, sparse and other variants," in *Learning in Graphical Models*, Eds. M. I. Jordan, Kluwer Academic Press, 1998.

[8] D. Tzikas et al., "The Variational Approximation for Bayesian Inference," *IEEE Signal Process. Mag.*, Nov. 2008.

[9] R. Shumway and D. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *Journal of Time Series Analysis*, 3(4), 253-265, 1982.

[10] J. Benesty et al., *Advances in Network and Acoustic Echo Cancellation*, Springer, Berlin, 2001.

[11] N. Goodman, "Statistical analysis based on a certain multivariate complex Gaussian distribution," *The Annals of Mathematical Statistics*, vol. 34, no. 1, pp. 152–177, March 1963.

[12] K. Petersen and M. Pedersen, *The Matrix Cookbook*, http://matrixcookbook.com, 2008.

[13] R. Shumway and D. Stoffer, *Time Series Analysis and Its Applications*, Springer, NY, USA, 2006.