Online Variational Message Passing in Hierarchical Autoregressive Models

Albert Podusenko and Wouter M. Kouw
Eindhoven University of Technology
Eindhoven, the Netherlands
{a.podusenko,w.m.kouw}@tue.nl

Bert de Vries
Eindhoven University of Technology, and GN Hearing
Eindhoven, the Netherlands
bert.de.vries@tue.nl

***Abstract*—Hierarchical autoregressive (AR) models can describe many complex physical processes. Unfortunately, online adaptation in these models under non-stationary conditions remains a challenge. In this paper, we track states and parameters in a hierarchical AR filter by means of variational message passing (VMP) in a factor graph. We derive VMP update rules for an "AR node" that can be re-used at various hierarchical levels and supports automated message passing-based inference for states and parameters. The proposed method is experimentally validated for a 2-level hierarchical AR model.**

I. Introduction

Autoregressive (AR) models predict future observations as a weighted combination of past observations. These models are extensively used to describe many natural processes [1]–[4]. Quite often, the statistics of dynamic processes vary over time, e.g. for speech signals. In order to capture time-varying dynamics, the AR model's coefficients should vary over time. An important variant for modeling time-varying AR coefficients is to let the dynamics of the AR coefficients themselves be modelled by an AR process (and so on), thus yielding a hierarchical AR model (HAR) [5] .

In this paper, we aim to solve the problem of Bayesian tracking of states and parameters in a hierarchical AR model. Unfortunately, the hierarchical structure of these models yields an inference problem that is not solvable in closed-form. Numerical approximation methods such as Monte Carlo sampling, are too slow for real-time inference in realistic models on small computing platforms.

Roberts and Penny [6] proposed a variational Bayes procedure for generalized autoregressive (GAR) models. Their work focused on parameter estimation of stationary signals with non-Gaussian and/or non-stationary noise processes. Here, we follow the variational lead by Roberts and Penny, but in order to develop a scalable and modular inference method that applies to tracking non-stationary signals as well, we employ variational message passing (VMP) on a factor graph. It lets us exploit the factorized (Markovian) structure of the HAR model, [7], [8].

In related work, [9] proposes a message-passing version of the expectation maximization (EM) algorithm to estimate AR coefficients. Unlike EM, which yields point estimates, VMP tracks approximate posterior distributions over the hidden states and parameters. Furthermore, we infer process noise precision, while [9] assumes that the noise is known.

The contributions in this paper include the following: First, we define an "AR node" for a Forney-style factor graph (FFG) and describe the factor graph structure of an HAR model that is composed of multiple AR nodes (Fig. 1 and Sec. II). Secondly, we derive new variational update rules for the AR nodes (Table I). With these rules, a hybrid message passing-based inference algorithm can be used to track time-varying coefficients and process noise parameters of the HAR model. Lastly, we experimentally validate our inference procedure by estimating coefficients and parameters in a 2-layer HAR model from a synthetic non-stationary data stream (Sec. III). Visualizations of the inferred coefficients in the upper layer of the HAR model show its ability to capture the time-varying dynamics.

II. Methods

A. Model specification

Consider a signal $y_t \in \mathbb{R}$ where t indexes discrete time steps. We write the dynamics of a 2-layer[1] autoregressive model (AR) for y_t as a state-space model as

$$\boldsymbol{\theta}_t^{(1)} = A(\boldsymbol{\theta}^{(2)})\boldsymbol{\theta}_{t-1}^{(1)} + \mathbf{c}v_t^{(1)} \tag{1a}$$

$$\boldsymbol{\theta}_t^{(0)} = A(\boldsymbol{\theta}_t^{(1)})\boldsymbol{\theta}_{t-1}^{(0)} + \mathbf{c}v_t^{(0)} \tag{1b}$$

$$y_t = \mathbf{c}^\top \boldsymbol{\theta}_t^{(0)} + w_t \tag{1c}$$

where $\mathbf{c} = (1, 0, ..., 0)^\top$, the ith layer state vector $\boldsymbol{\theta}_t^{(i)} = (\theta_t^{(i)}, \theta_{t-1}^{(i)}, \ldots, \theta_{t-M+1}^{(i)})^\top$, and $v_t^{(i)}$ and w_t represent zero-mean Gaussian noise signals. The state transition matrix $A(\boldsymbol{\theta})$ is given by

$$A(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{\theta}^\top \\ \mathbf{I}_{M-1} \quad \mathbf{0} \end{bmatrix}, \quad \mathbf{I}_{M-1} = \begin{bmatrix} 1_1 & 0 & \ldots & 0 \\ 0 & 1_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1_{M-1} \end{bmatrix}.$$

In this model, Eq. 1b is a regular AR model in state space form for signal $\boldsymbol{\theta}_t^{(0)}$, parameterized by $\boldsymbol{\theta}_t^{(1)}$, which by themselves are dynamically generated by AR model Eq. 1a. Eq. 1c is an observation model that selects the first component of state vector $\boldsymbol{\theta}_t^{(0)}$ and adds observation noise.

[1]The description generalizes easily. We use a 2-layer model for simplicity.

In this paper, we will develop inference methods on this model by message passing on a Forney-style Factor Graph (FFG) representation of the model. Factor graphs are graphical models for factorized probability distributions and the hierarchical AR model can alternatively be written as the following (factorized) *generative probabilistic distribution*:

$$p(\Theta, \mathbf{y}, \boldsymbol{\gamma}) = \tag{2a}$$

$$\underbrace{p(\Theta_0)p(\boldsymbol{\gamma})}_{\text{priors}} \prod_{t=1}^{T} \underbrace{p(y_t|\boldsymbol{\theta}_t^{(0)})}_{\text{observation}} \prod_{i=0}^{N} \underbrace{p(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_t^{(i+1)}, \boldsymbol{\theta}_{t-1}^{(i)}, \gamma^{(i)})}_{\text{state transition}}$$

$$p(\Theta_0) = \prod_{i=0}^{N} \mathcal{N}(\boldsymbol{\theta}_0^{(i)}|m_{\boldsymbol{\theta}_0}^{(i)}, V_{\boldsymbol{\theta}_0}^{(i)}) \tag{2b}$$

$$p(\boldsymbol{\gamma}) = \prod_{i=0}^{N} \Gamma(\gamma^{(i)}|\alpha^{(i)}, \beta^{(i)}) \tag{2c}$$

$$p(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_t^{(i+1)}, \boldsymbol{\theta}_{t-1}^{(i)}, \gamma^{(i)}) = \mathcal{N}\Big(\boldsymbol{\theta}_t^{(i)}|A(\boldsymbol{\theta}_t^{(i+1)})\boldsymbol{\theta}_{t-1}^{(i)}, V^{(i)}\Big) \tag{2d}$$

$$p(y_t|\boldsymbol{\theta}_t^{(0)}) = \mathcal{N}\Big(y_t|, \mathbf{c}^\top \boldsymbol{\theta}_t^{(0)}, \tau^{-1}\Big) \tag{2e}$$

where $\mathbf{y} \triangleq y_{1:T} = (y_1, y_2, ..., y_T)$ is a sequence of observations and $\tau \in \mathbb{R}^+$ is a precision parameter for the Gaussian observation noise. Eqs. 2b and 2c are priors for the hidden states $\Theta_t = (\boldsymbol{\theta}_t^{(0)}, \boldsymbol{\theta}_t^{(1)}, \ldots, \boldsymbol{\theta}_t^{(N)})$ and parameters $\boldsymbol{\gamma} = (\gamma^{(0)}, \gamma^{(1)}, \ldots, \gamma^{(N)})$, respectively. The covariance matrix $V^{(i)}$ is defined as

$$V^{(i)} = \begin{bmatrix} 1/\gamma^{(i)} & 0 & \ldots & 0 \\ 0 & 0 & \ldots & \vdots \\ \vdots & & \ddots & \end{bmatrix} \tag{3}$$

The generative model introduced in Eq. 2 can be visually represented by a Forney-style factor graph (FFG) as shown in Fig. 1 (the details of factor graphs will be discussed in Sec. II-C).

B. Problem: online inference of states and parameters

The central quantity of interest is the (joint) posterior distribution of states Θ_t and parameters $\boldsymbol{\gamma}$, given all past observations $\mathbf{y}_{1:t}$. In principle, this posterior can be obtained through recursive application of Bayes rule as

$$\underbrace{p(\Theta_t, \gamma|\mathbf{y}_{1:t})}_{\text{posterior}} = \frac{1}{\underbrace{p(y_t|\mathbf{y}_{1:t-1})}_{\text{evidence}}} \underbrace{p(y_t|\Theta_t)}_{\text{likelihood}}$$

$$\cdot \int \underbrace{p(\Theta_t|\Theta_{t-1}, \boldsymbol{\gamma})}_{\text{state transitions}} \underbrace{p(\Theta_{t-1}, \gamma|\mathbf{y}_{1:t-1})}_{\text{prior}} \mathrm{d}\Theta_{t-1} \tag{4}$$

Unfortunately, this expression is analytically intractable due to the integration over large state spaces and non-conjugate prior-posterior pairings. Moreover, evaluation of the evidence factor involves an integral without a closed-form solution. In this paper we work out an approximate inference solution based on variational message passing (VMP).

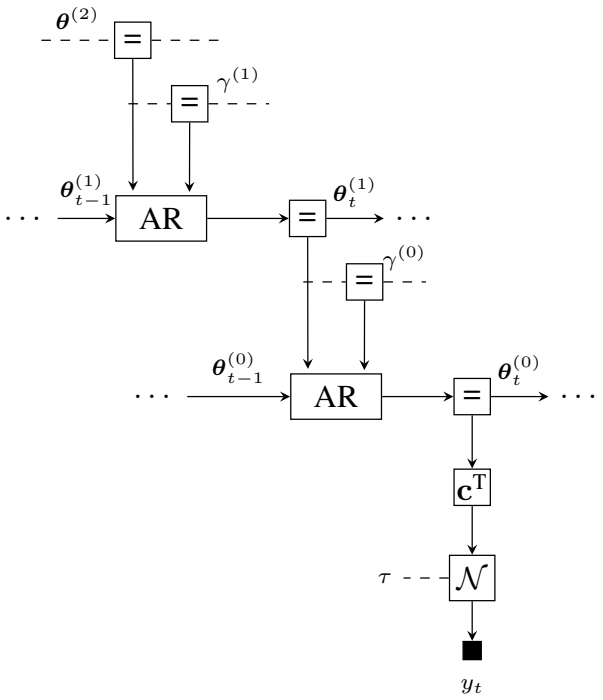


Fig. 1. One time segment of a Forney-style factor graph (FFG) for the 2- layer HAR model as defined by Eq. 2. The AR node denotes a transition model (Eq. 2d). The small black node corresponds to an observed variable (y_t); medium-sized nodes represent deterministic factors and large nodes denote stochastic factors. Solid and dashed edges are associated with states and parameters respectively. The dotted edges on the left and right of the graph indicate that this model extends in the same way to the other time steps. The arrowheads indicate the "generative" direction but do not affect any inference computations.

C. Forney-style Factor Graphs

In this section, we shortly summarize Forney-style Factor Graphs (FFG). An FFG is a diagram of a factorization of a function of several variables, where variables and factors are represented by edges (or half-edges) and nodes respectively [10]. An edge is connected to a node if and only if the (edge) variable is an argument of the (node) function. If a variable appears in more than two factors, equality ("branching") nodes connect copies of the variable to the other factors under the constraint that the marginal beliefs for all copies (and original variable) are equal.

As an example, consider the factorized probabilistic model

$$p(x_1, x_2, x_3, x_4) = f_A(x_1, x_2) f_B(x_2, x_3) f_C(x_3, x_4)\,, \tag{5}$$

and assume that we are interested in the marginal distribution of x_3, given by

$$p(x_3) = \int\int\int p(x_1, x_2, x_3, x_4)\,\mathrm{d}x_1 \mathrm{d}x_2 \mathrm{d}x_4 \tag{6}$$

Due to the factorization of Eq. 5, we can optimize the amount of computations by using the distributive law (i.e., by moving

factors over the integration signs), leading to:

$$p(x_3) = \underbrace{\underbrace{\int f_A(x_1,x_2)\mathrm{d}x_1}_{\overrightarrow{\mu}(x_2)} \int f_B(x_2,x_3)\mathrm{d}x_2}_{\overrightarrow{\mu}(x_3)} \underbrace{\int f_C(x_4,x_3)\mathrm{d}x_4}_{\overleftarrow{\mu}(x_3)} \tag{7}$$

In this way, the three-dimensional integral of Eq. 6 reduces to (multiplications of the results of) much simpler one-dimensional integrals. In the context of an FFG, the results of these sub-integrals can be interpreted as locally computable messages as depicted in Fig. 2.

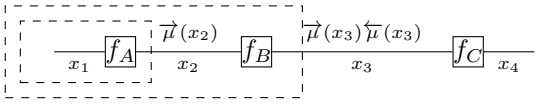


Fig. 2. An FFG corresponding to the model given by Eq. 5, including messages as per Eq. 7.

This example illustrates the idea of the *sum-product rule* which states that for a generic node $f(y, x_1, \ldots, x_n)$, the outgoing message $\overrightarrow{\mu}(y)$ is given by

$$\overrightarrow{\mu}(y) = \int \cdots \int f(y, x_1, \ldots, x_n) \prod_{i=1}^{n} \overrightarrow{\mu}(x_i)\mathrm{d}x_i \,. \tag{8}$$

A more detailed explanation of sum-product message passing in FFGs can be found in [11].

D. Variational Message Passing

Since the integrals in Eq. 4 are not tractable, we cannot solve the inference problem by only using sum-product messages. Therefore we resort to Variational Message Passing (VMP), which is an *approximate* Bayesian inference technique based on minimization of the variational free energy (FE),

$$F_t[q] \triangleq \int q(\Theta_t,\gamma) \log \frac{q(\Theta_t,\gamma)}{p(y_t,\Theta_t,\gamma|\mathbf{y}_{1:t-1})} \mathrm{d}\Theta_t \mathrm{d}\gamma \tag{9}$$

$$= \underbrace{-\log p(y_t|\mathbf{y}_{1:t-1})}_{-\text{log evidence}} + \underbrace{\int q(\Theta_t,\gamma) \log \frac{q(\Theta_t,\gamma)}{p(\Theta_t,\gamma|\mathbf{y}_{1:t})} \mathrm{d}\Theta_t \mathrm{d}\gamma}_{\text{KL divergence}}$$

where $q(\Theta_t, \gamma)$ is an approximate posterior distribution (also called the *recognition distribution*) for the hidden variables Θ_t and γ. Since the Kullback-Leibler (KL) divergence is guaranteed to be non-negative and only equals zero if $q(\Theta_t, \gamma) = p(\Theta_t, \gamma|\mathbf{y}_{1:t})$, minimization of $F_t[q]$ with respect to q leads to $q(\Theta_t, \gamma) \approx p(\Theta_t, \gamma|\mathbf{y}_{1:t})$ and $F_t[q] \approx -\log p(y_t|\mathbf{y}_{1:t-1})$. Thus, minimization of Eq. 9 approximately solves the Bayesian filtering problem of Eq. 4, [12].

Eq. 9 can be minimized by a "variational" message passing algorithm [7], [8]. Consider a generic node $f(y, x_1, \ldots, x_n)$ as depicted in Fig. 3. It can be shown that minimization of FE results in sending a (variational) message of the form

$$\overrightarrow{\nu}(y) \propto \exp\left(\int q(\mathbf{x}) \log f(y, x_1, \ldots, x_n)\mathrm{d}\mathbf{x}\right), \tag{10}$$

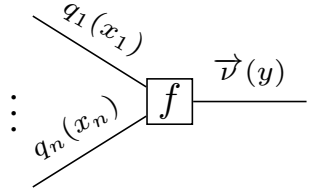


Fig. 3. A generic node $f(y, x_1, ..., x_n)$ with incoming variational messages $q_i(x_i)$ and outgoing variational message $\overrightarrow{\nu}(y)$, see Eq. 10.

where $\mathbf{x} = (x_1, \ldots, x_n)$ and $q(\mathbf{x}) = \prod q_i(x_i)$, [7]. The approximate marginal $q(y)$ can be obtained by multiplying incoming and outgoing messages on the edge for y, i.e.,

$$q(y) = \overrightarrow{\nu}(y)\overleftarrow{\nu}(y) \,. \tag{11}$$

For a more detailed explanation of VMP in FFGs, we refer to [8].

E. Online VMP for HAR models

Within the context of FFGs, a hierarchical autoregressive (HAR) model is a configuration of stacked AR nodes, see Fig. 4. An AR node is internally structured as shown in the top panel of Table I. We are interested in tracking states and parameters by message passing. While a sum-product message from the observation block (13) is possible, the outgoing sum-product messages from the AR nodes (e.g. (8), (14), (15) for the first layer) do not have a closed-form solution. Hence, we use a hybrid message passing scheme consisting of sum-product messages when possible and otherwise we use variational messages. Due to the modularity of the FFG framework, we only need to work out the message update rules for an AR node once and re-use these rules at all instances of the AR node.

For the sake of notational generality, we now replace $\boldsymbol{\theta}_t^{(i)}$, $\boldsymbol{\theta}_{t-1}^{(i)}$, $\boldsymbol{\theta}_t^{(i+1)}$ and $\gamma^{(i)}$ by $\mathbf{y}$, $\mathbf{x}$, $\boldsymbol{\theta}$ and γ respectively. We then specify the AR node function by the factor

$$f_{AR}(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}, \gamma) = \mathcal{N}\left(\mathbf{y} \,|\, A(\boldsymbol{\theta})\mathbf{x}, V\right) \,. \tag{12}$$

with V as defined by Eq. 3. In Table I, we provide the full set of variational messages that we derived using the naive mean field assumption[2] over $\mathbf{y}, \mathbf{x}, \theta, \gamma$, i.e., we assumed that

$$q(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}, \gamma) = q(\mathbf{y})q(\mathbf{x})q(\boldsymbol{\theta})q(\gamma) \,. \tag{13}$$

These update rules support automated message passing-based online inference in complex models with AR nodes as sub-models, see Fig. 4.

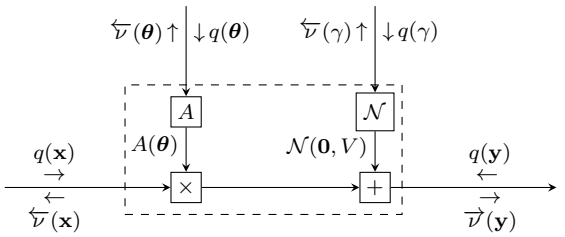
III. Experimental validation

In order to validate inference with the tabulated AR-node messages in a full HAR(1) model, we modeled a data set with a 2-layer AR model. The data set was generated by the 2-layer AR model in Eq. 2, with $\tau = 2$, $\gamma^{(0)} = 1.0$ and $\gamma^{(1)} = 2$,

[2]Derivations of update rules can be found at http://biaslab.github.io/pdf/isit2020/a_podusenko_AR_meanfield_derivations.pdf.

TABLE I

VARIATIONAL MESSAGE UPDATE RULES FOR THE AUTOREGRESSIVE (AR) NODE (DASHED BOX) OF EQ. 12. DISTRIBUTIONS $q(\boldsymbol{\theta}) = \mathcal{N}(m_{\boldsymbol{\theta}}, V_{\boldsymbol{\theta}})$, $q(\mathbf{x}) = \mathcal{N}(m_{\mathbf{x}}, V_{\mathbf{x}})$, $q(\mathbf{y}) = \mathcal{N}(m_{\mathbf{y}}, V_{\mathbf{y}})$ AND $q(\gamma) = \Gamma(\alpha, \beta)$ ARE ASSOCIATED WITH INCOMING MESSAGES. OUTGOING MESSAGES $\nu(\cdot)$ ARE TABULATED BELOW. THE SUPERSCRIPTS IN $V_y^{(1,1)}$ AND $m_{\mathbf{y}}^{(1)}$ DENOTE THE FIRST ELEMENT OF THE MATRIX AND VECTOR, RESPECTIVELY.

Node	
	
Messages	**Update Rule**
$\overrightarrow{\nu}(\mathbf{y})$	$\mathcal{N}\left(A(m_{\boldsymbol{\theta}})m_{\mathbf{x}}, m_W^{-1}\right)$
$\overleftarrow{\nu}(\mathbf{x})$	$\mathcal{N}\left(\mathbf{D}_1^{-1}\mathbf{z}_1, \mathbf{D}_1^{-1}\right)$
$\overleftarrow{\nu}(\boldsymbol{\theta})$	$\mathcal{N}\left(\mathbf{D}_2^{-1}\mathbf{z}_2, \mathbf{D}_2^{-1}\right)$
$\overleftarrow{\nu}(\gamma)$	$\Gamma\left(\frac{3}{2}, \frac{B}{2}\right)$
Auxiliary variables	

$$\mathbf{D}_1 = A(m_{\boldsymbol{\theta}})^\top m_W A(m_{\boldsymbol{\theta}}) + V_{\boldsymbol{\theta}} m_\gamma$$
$$\mathbf{z}_1 = A(m_{\boldsymbol{\theta}})^\top m_W m_{\mathbf{y}}$$
$$\mathbf{D}_2 = V_{\mathbf{x}} m_\gamma + m_{\mathbf{x}} m_\gamma m_{\mathbf{x}}^{\mathrm{T}}$$
$$\mathbf{z}_2 = m_{\mathbf{x}} \mathbf{c}^\top m_W m_{\mathbf{y}}$$
$$B = V_{\mathbf{y}}^{(1,1)} + m_{\mathbf{y}}^{(1)} m_{\mathbf{y}}^{(1)} - 2 m_{\mathbf{y}}^{(1)} m_{\boldsymbol{\theta}}^\top m_{\mathbf{x}} + m_{\mathbf{x}}^\top V_{\boldsymbol{\theta}} m_{\mathbf{x}} + m_{\boldsymbol{\theta}}^\top (V_{\mathbf{x}} + m_{\mathbf{x}} m_{\mathbf{x}}^\top) m_{\boldsymbol{\theta}}$$
$$m_W = \begin{bmatrix} m_\gamma & 0 & \dots & 0 \\ 0 & 1/\epsilon & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1/\epsilon \end{bmatrix}$$
$$\epsilon > 0\,, \quad m_\gamma = \frac{\alpha}{\beta}$$

$\boldsymbol{\theta}^{(2)} = -0.556$. The inference schedule for one time segment for the HAR(1) model is depicted in Fig. 4. Messages ③ and ⑦ carry posterior estimates from the previous time step $t-1$. The messages ⑫ and ⑯ propagate estimates $q(\boldsymbol{\theta}_t^{(0)})$ and $q(\boldsymbol{\theta}_t^{(1)})$ of the state posteriors $p(\boldsymbol{\theta}_t^{(0)}|\mathbf{y}_{1:t})$ and $p(\boldsymbol{\theta}_t^{(1)}|\mathbf{y}_{1:t})$ respectively. The messages ⑭ and ⑮ carry the parameter estimates $q(\gamma^{(0)})$ for $p(\gamma^{(0)}|\mathbf{y}_{1:t})$ and $q(\gamma^{(1)})$ for $p(\gamma^{(1)}|\mathbf{y}_{1:t})$. The AR node sends variational messages according to the update rules described in Table I. To update the posteriors we iterate through messages ①-⑲ for each time segment. We implemented our method with the open source Julia package `ForneyLab` that is under development in our research group [13].

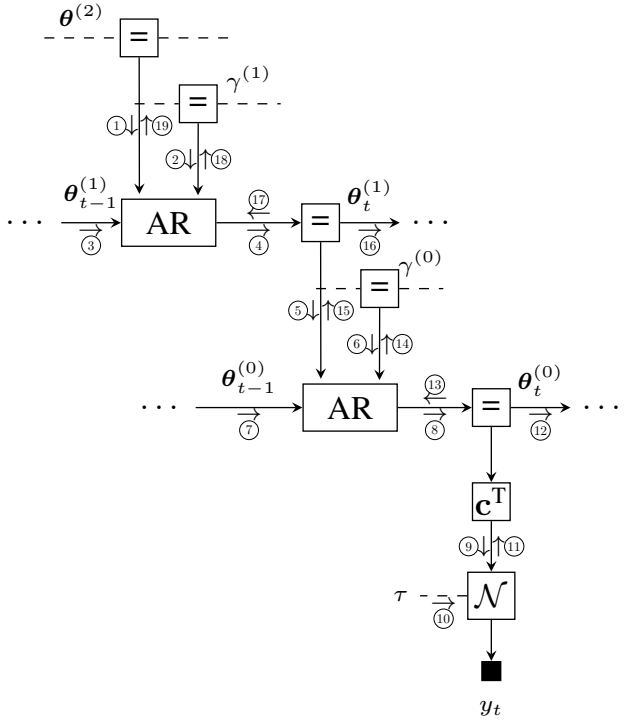


Fig. 4. The message passing schedule for online state estimation in a 2-layer HAR(M). The messages are computed in increasing number order.

A. Baseline models

We compared the performance of the 2-layer HAR(1) model to an AR(1) model and to a random walk (RW) model. The AR(1) model is a special case of HAR(1) in the sense that it has only one layer,

$$p(\Theta, \mathbf{y}, \gamma) = p(\boldsymbol{\theta}_0^{(0)}) p(\gamma^{(0)}) \prod_{t=1}^{T} p(y_t|\boldsymbol{\theta}_t^{(0)}) p(\boldsymbol{\theta}_t^{(0)}|\boldsymbol{\theta}_{t-1}^{(0)}, \boldsymbol{\theta}^{(1)}, \gamma^{(0)}) \quad (14)$$

and the RW model is a special case of AR(1) in that the coefficient is fixed at $\boldsymbol{\theta}^{(1)} = 1$. Therefore, comparing RW with AR(1) shows the effect of freeing the autoregressive coefficient, while comparing AR(1) to HAR(1) shows the effect of allowing the autoregressive coefficient to be time-varying. Note that we used identical prior parameters for all three models where possible, to keep the comparison fair.[3]

B. Performance metric

In order to assess model performance, we track a loss function

$$\mathcal{L}_t(m_{\theta_t^{(0)}}, v_{\theta_t^{(0)}}, \theta_t^{(0)}) = \frac{(m_{\theta_t^{(0)}} - \theta_t^{(0)})^2}{v_{\theta_t^{(0)}}} + \log v_{\theta_t^{(0)}}$$

[3]The Jupyter notebook with the experiments can be found at https://github.com/biaslab/ISIT-2020

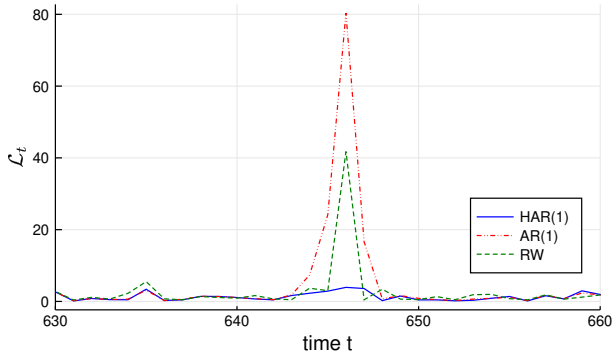


Fig. 5. Zoom-in of performance signal $\mathcal{L}_t$ values from $t = 630$ to $t = 660$ for the validation experiment in Sec. III. Lower values for $\mathcal{L}_t$ correspond to better performance.

where $m_{\theta_t^{(0)}}$ and $v_{\theta_t^{(0)}}$ are estimated mean and variance of the hidden state, while $\theta_t^{(0)}$ corresponds to the real value (used during data synthesis).

This metric is inspired by the Free Energy functional (energy minus entropy), where both weighted prediction errors *and* uncertainty (large variance) are penalized. When comparing models, lower values of $\mathcal{L} = (1/T)\sum_{t=1}^{T}\mathcal{L}_t$ signify better performance. We expect the HAR model to exhibit the best performance as it is the closest match to the data generating process.

C. Results

Fig. 6 plots state estimates for HAR(1), AR(1) and RW from $t = 600$ to $t = 700$. At $t \approx 645$, the true signal spikes (red line in the top sub-figure). The HAR(1) model captures this in the state of its top layer $\boldsymbol{\theta}^{(1)}$ (black dotted line in top sub-figure). The result is that HAR(1) approaches the true state in the lower layer $\boldsymbol{\theta}^{(0)}$ (black dotted line approaches pink solid line around $t \approx 645$ in second sub-figure), which is something that RW and AR(1) fails at (third and fourth sub-figures). Note that the RW model approaches the true state around the top of the spike, but falls short of the bottom of the spike. In general, the RW model is less accurate since the model is too simple. Table II reports the performance scores for the three models over the entire time series (T=1000), i.e. $\mathcal{L}$. Evidently, HAR(1) performs (on average) better than AR(1), which in turn outperforms the RW model. In Fig. 5 we plot the values of $\mathcal{L}_t$ for 30 time steps. Both RW and AR(1) models clearly fail at the spike around $t \approx 645$.

TABLE II
PERFORMANCE SCORES ($\mathcal{L}$, ROUNDED TO THE SECODN DECIMAL POINT) FOR THE HIERARCHICAL AUTOREGRESSIVE (HAR) MODEL, THE AUTOREGRESSIVE MODEL (AR) AND THE RANDOM WALK (RW), AVERAGED OVER THE FULL LENGTH OF THE SIGNAL ($t = 1$ TO 1000).

	HAR	AR	RW
$\mathcal{L}$	1.08	1.46	1.49

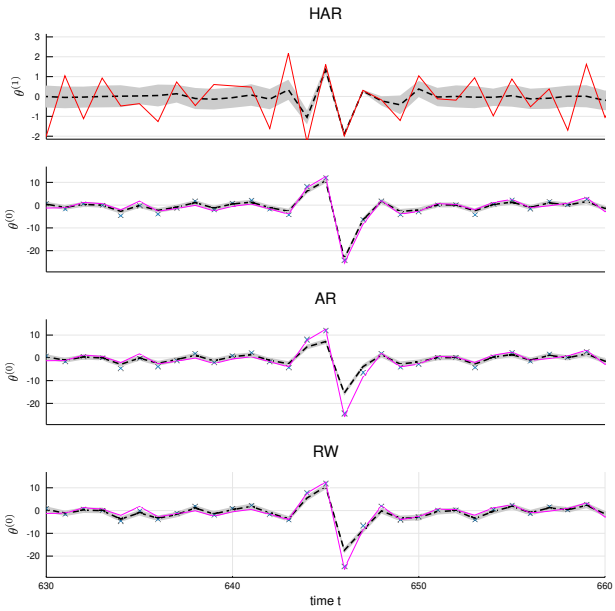


Fig. 6. Simulation results. The dashed line corresponds to the expected mean value of the posterior estimates for hidden states. The shadowed region corresponds to one standard deviation below and above the mean. The top two graphs show inferred states for the first and second layers, as recovered by the HAR(1) model. The two bottom plots display AR and RW inference results.

IV. CONCLUSIONS

In this paper, we presented a hierarchical autoregressive model and showed how to track the states and parameters by automatable message passing-based inference in a factor graph framework. We derived variational message passing update rules for an "AR node" that can be applied locally wherever AR sub-models appear in a more complex model. In the future, we plan to extend our investigations to online tracking of time-varying process noise statistics by message passing.

ACKNOWLEDGEMENTS

This work was partly financed by research programme ZERO with project number P15-06, which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO).

REFERENCES

[1] R. Morris, M. Clements, and J. Collura, "Autoregressive parameter estimation of speech in noise," in *IEEE Workshop on Speech Coding*, 2002, pp. 181–183.
[2] M. A. Berezina, D. Rudoy, and P. J. Wolfe, "Autoregressive modeling of voiced speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 5042–5045.
[3] C. H. You, S. Rahardja, and S. N. Koh, "Autoregressive Parameter Estimation for Kalman Filtering Speech Enhancement," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, vol. 4, Apr. 2007, pp. IV–913–IV–916, iSSN: 2379-190X.
[4] M. Shannon, H. Zen, and W. Byrne, "Autoregressive Models for Statistical Parametric Speech Synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 3, pp. 587–597, Mar. 2013.
[5] O. Kakusho and M. Yanagida, "Hierarchical AR model for time varying speech signals," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 7, 1982, pp. 1295–1298.

[6] S. J. Roberts and W. D. Penny, "Variational Bayes for generalized autoregressive models," *IEEE Transactions on Signal Processing*, vol. 50, no. 9, pp. 2245–2257, Sep. 2002.

[7] J. Winn and C. M. Bishop, "Variational message passing," *Journal of Machine Learning Research*, vol. 6, no. Apr, pp. 661–694, 2005. [Online]. Available: http://www.jmlr.org/papers/volume6/winn05a/winn05a.pdf

[8] J. Dauwels, "On Variational Message Passing on Factor Graphs," in *IEEE International Symposium on Information Theory*, Jun. 2007, pp. 2546–2550. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/4557602

[9] J. Dauwels, A. Eckford, S. Korl, and H.-A. Loeliger, "Expectation maximization as message passing-part I: Principles and gaussian messages," *arXiv preprint arXiv:0910.2832*, 2009. [Online]. Available: http://arxiv.org/abs/0910.2832

[10] H.-A. Loeliger, "An introduction to factor graphs," *Signal Processing Magazine, IEEE*, vol. 21, no. 1, pp. 28–41, 2004. [Online]. Available: https://ieeexplore.ieee.org/document/1267047

[11] S. Korl, "A factor graph approach to signal modelling, system identification and filtering," Ph.D. dissertation, Swiss Federal Institute of Technology, Zurich, 2005.

[12] D. Barber, *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.

[13] M. Cox, T. van de Laar, and B. de Vries, "A factor graph approach to automated design of Bayesian signal processing algorithms," *International Journal of Approximate Reasoning*, vol. 104, pp. 185–204, 2019.