

Message Passing-based Inference in Switching Autoregressive Models

Albert Podusenko Bart van Erp Dmitry Bagaev İsmail Şenöz Bert de Vries
Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

Abstract—The switching autoregressive model is a flexible model for signals generated by non-stationary processes. Unfortunately, evaluation of the exact posterior distributions of the latent variables for a switching autoregressive model is analytically intractable, and this limits the applicability of switching autoregressive models in practical signal processing tasks. In this paper we present a message passing-based approach for computing approximate posterior distributions in the switching autoregressive model. Our solution tracks approximate posterior distributions in a modular way and easily extends to more complicated model variations. The proposed message passing algorithm is verified and validated on synthetic and acoustic data sets respectively.

Index Terms—Message Passing, State Estimation, Switching Autoregressive Models, Variational Inference

I. INTRODUCTION

Autoregressive (AR) models have been widely used to represent acoustic signals, such as speech signals [1], [2] or background noise [3], [4]. In order to take into account non-stationary behaviour, switching autoregressive (SwAR) models have been developed as an extension to standard AR models [5], [6, Ch. 24.6]. These models allow for representing individual phonemes in speech or different types of background noise. Besides, SwAR models have been successfully applied to other tasks, including the detection of influenza epidemics [7], heart sound segmentation [8] and wind time series processing [9]. This extension from the original AR model may lead to increasing model performance, but also leads to a more complicated inference procedure.

Technically, the SwAR model only differs from the regular AR model through its prior distributions on the parameters, as will be specified in detail in Section II. Instead of deriving all update equations for state and parameter estimation in this specific model by hand, as was done for the simplified model in [6, Ch. 24.6], we automate inference by message passing in (Forney-style) factor graph (FFG) representation of the model [10], [11]. The local message update equations have been pre-derived for the constituent factor nodes of the SwAR model in earlier works [12, Apps. 2 & 9], [13], which allows us to automatically generate an inference algorithm for the SwAR model.

This paper describes a message passing-based approach for performing probabilistic inference in the switching autoregressive model. We make the following contributions:

- A switching auto-regressive model is specified where both states and parameters are treated as latent variables in Section II.
- The basic SwAR model is extended with temporal dynamics for the active switching states evolving over a different time scale in Section II.
- We state our problem definition as an inference task on the SwAR model in Section III, and show how this inference task can be realized through message passing-based inference in an FFG in Section IV.
- We demonstrate our proposed methodology through a set of verification and validation experiments in Section V.

Finally, we discuss the obtained results and conclude the paper in Section VI.

II. MODEL SPECIFICATION

Let $\mathbf{y}_t \triangleq [y_t, \dots, y_{t-N+1}]^\top \in \mathbb{R}^N$, denote a vector of the N latest observations at time t . The likelihood function of an SwAR model is defined as

$$y_t \sim \mathcal{N}(\boldsymbol{\theta}_k^\top \mathbf{y}_{t-1}, \gamma_k), \quad (1)$$

where we use $\mathcal{N}(\mu, \Lambda)$ to denote a Gaussian distribution with mean μ and precision Λ . $\boldsymbol{\theta}_k = [\theta_{1k}, \dots, \theta_{Nk}]^\top \in \mathbb{R}^N$ and $\gamma_k \in \mathbb{R}_{>0}$ denote the autoregressive coefficients and process noise precision of the N^{th} -order SwAR model, respectively.

The vector of previous observations \mathbf{y}_{t-1} is updated with the next observation y_t according to [14] by

$$\mathbf{y}_t = \mathbf{S}\mathbf{y}_{t-1} + \mathbf{c}y_t \quad (2)$$

where

$$\mathbf{S} \triangleq \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{I}_{M-1} & \mathbf{0} \end{bmatrix}, \quad \mathbf{c} \triangleq [1, 0, \dots, 0]^\top. \quad (3)$$

We assume the parameters of SwAR to be stationary over longer segments of time and therefore index them with the slower-evolving switching state index $k = 1, \dots, K$, related to t as $k = \lceil t/W \rceil$. Here, $\lceil \cdot \rceil$ denotes the ceiling function that returns the largest integer smaller or equal than its argument, while W is the window length. The above equation makes sure that k is intuitively aligned with segments of length W , i.e. $t \in [1, W]$ corresponds to $k = 1$. To denote the start and end indices of the time segment corresponding to switching state index k , we define $t^- = (k-1)W + 1$ and $t^+ = kW$ as an implicit function of k , respectively. Implicitly we also

Correspondence: a.podusenko@tue.nl

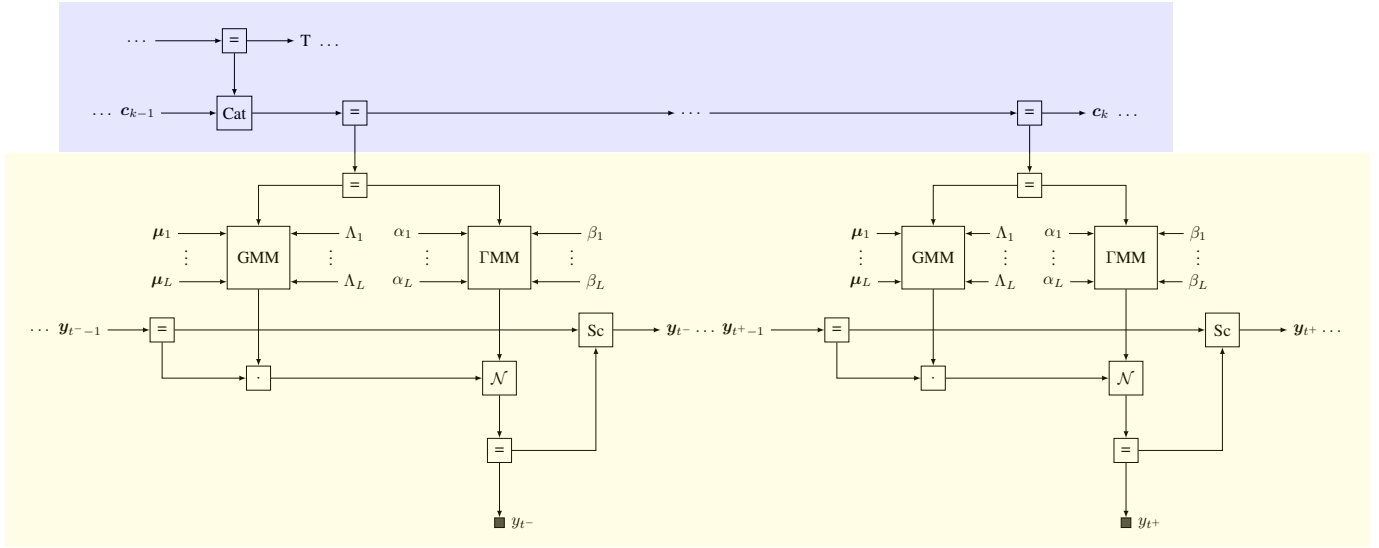


Fig. 1. A Forney-style factor graph of a single switching state time slice of the switching autoregressive model. Each switching state time segment indexed over k (blue part) is connected to a total of W repetitions of the part of the model indexed over t (yellow part). The GMM and Γ M nodes denote the Gaussian and Gamma mixture nodes respectively. The Sc node represents (2). Horizontal dots indicate the extension of the graph in the same way to the other time steps. Vertical dots encapsulate mixture components from 1 to L . The equality node = imposes a constraint on the marginals of variables associated with adjacent edges to be equal. Propagating messages backwards from future time steps results in an inference smoothing algorithm, whereas if we only propagate messages forward in time, a filtering algorithm results.

constrain (1) to only be valid for matching time indices, i.e. for $t = t^-, t^- + 1, \dots, t^+$.

The AR likelihood function of (1) is extended with the mixture models

$$\theta_k \sim \prod_{l=1}^L \mathcal{N}(\mu_l, \Lambda_l)^{c_{lk}} \quad \gamma_k \sim \prod_{l=1}^L \Gamma(\alpha_l, \beta_l)^{c_{lk}} \quad (4)$$

to form a SwAR model with L switching states or contexts. Here $\Gamma(\alpha, \beta)$ denotes the Gamma distribution with shape and rate parameters α and β , respectively. The variable $c_k = [c_{1k}, \dots, c_{Lk}]^T$ denotes a 1-of- L binary vector with elements $c_{lk} \in \{0, 1\}$, constrained by $\sum_l c_{lk} = 1$. The switching behaviour is achieved by modeling the temporal dynamics as

$$c_k \sim \text{Cat}(Tc_{k-1}), \quad (5)$$

where $\text{Cat}(\pi)$ denotes a categorical distribution with event probabilities π . We model the individual columns of the transition matrix T by a Dirichlet distribution $\text{Dir}(\zeta)$ as

$$T_{1:L,j} \sim \text{Dir}(\zeta_j), \quad (6)$$

where ζ_j denotes the vector of concentration parameters corresponding to the j^{th} column of T . The switching state is initialized by a categorical distribution as

$$c_0 \sim \text{Cat}(\pi_0) = \prod_{l=1}^L \pi_{l0}^{c_{l0}} \quad \text{such that} \quad \sum_{l=1}^L \pi_{l0} = 1, \quad (7)$$

where the individual event probabilities can be chosen as $\pi_{l0} = 1/L$ if the initial switching state is unknown. Additionally, we

assign prior probability distributions to the hyperparameters of the SwAR model in (4):

$$\begin{aligned} \mu_l &\sim \mathcal{N}(\mathbf{m}_{l0}, \mathbf{v}_{l0}) & \Lambda_l &\sim \mathcal{W}(\mathbf{V}_0, n_0) \\ \alpha_l &\sim \Gamma(a_0^{\alpha_l}, b_0^{\alpha_l}) & \beta_l &\sim \Gamma(a_0^{\beta_l}, b_0^{\beta_l}) \end{aligned} \quad (8)$$

with $\mathcal{W}(\cdot, \cdot)$ denoting the Wishart distribution.

The SwAR model described by (1)-(8) can be represented by a Forney-style Factor Graph (FFG) as depicted in Figure 1. An FFG is an *undirected* graph where nodes represent factors of a global function and edges represent variables [10]. In an FFG, an edge is connected to a node if and only if the factor corresponding to the node is a function of the variable corresponding to the edge. If the variable is shared between more than 2 factors, we can make use of equality nodes of type $\delta(x - x')\delta(x - x'')$ that constrain the beliefs over two ‘‘copy variables’’ x' and x'' to be equal to the belief over x [15]. In an FFG, factors are drawn as square open nodes and observations or fixed variables are represented by small black squares, whose factors can be regarded as Dirac delta functions centered on the observed value. For a detailed explanation of the FFG formalism, we refer to [11], [16], [17].

III. PROBLEM STATEMENT

Given an SwAR model and a collection of observations \mathbf{y} , we are interested in tracking the marginal distributions of the model’s latent variables. Computation of these posterior distributions requires the integration and summation of the model specified by (1)-(6) with respect to all nuisance variables. These computations do not yield any analytical solutions and therefore lead to intractable probabilistic inference. This paper addresses the problem of computing approximate marginal distributions in the SwAR model.

IV. METHODS

In this section we describe how probabilistic inference can be realized in the SwAR model.

A. Variational message passing

The factorized structure of the SwAR model allows for the distributed calculation of the posterior distributions of its variables through a set of smaller local computations called messages. Intractability in these computations prevents us from performing exact message passing-based inference, also known as belief propagation [18] or the sum-product algorithm [19]. Consequently we resort to variational message passing (VMP) [20], [21].

To illustrate this, consider the probabilistic model $p(\mathbf{y}, \mathbf{z})$, with observations \mathbf{y} and latent variables \mathbf{z} . As the computation of the exact posterior $p(\mathbf{z}|\mathbf{y})$ is intractable, we resort to variational inference, where we approximate the true posterior distribution by the tractable approximate posterior distribution $q(\mathbf{z}) \approx p(\mathbf{z}|\mathbf{y})$. Probabilistic inference then concerns the minimization of the variational free energy (VFE) functional

$$F[q] = D_{\text{KL}}[q(\mathbf{z})||p(\mathbf{z}|\mathbf{y})] - \ln p(\mathbf{y}), \quad (9)$$

where D_{KL} is the Kullback-Leibler divergence. To enable efficient optimisation of the VFE for the SwAR model we assume an additional factorisation on $q(\mathbf{z})$,

$$q(\mathbf{z}) = \prod_a q_a(\mathbf{z}_a), \quad (10)$$

where \mathbf{z}_a refers to a set of node-bound local variables (one or many) such that $\cup_a \mathbf{z}_a = \mathbf{z}$. VMP concerns the iterative updating of marginals as $q_j(\mathbf{z}_j) \propto \vec{v}(z_j) \cdot \bar{v}(z_j)$, where $\vec{v}(z_j)$ and $\bar{v}(z_j)$ are forward and backward variational messages on edge z_j . The outgoing variational message $\bar{v}(z_j)$ on edge z_j from a factor $f(\mathbf{z})$, with incoming marginals $q_i(z_i)$ for $i \neq j$, can be derived as [21]

$$\bar{v}(z_j) \propto \exp \int \prod_{i \neq j} q_i(z_i) f(\mathbf{z}) d\mathbf{z}_{\setminus j}. \quad (11)$$

The approximate marginals $q_i(z_i)$ and variational messages $\vec{v}(z_j)$ and $\bar{v}(z_j)$ are iteratively updated until the VFE converges.

B. Expectation maximization

As a further specification of the VMP procedure, we can constrain the form of approximate marginals to $q_j(z_j) = \delta(z_j - \hat{z}_j)$. By selecting \hat{z}_j through the optimization problem

$$\hat{z}_j = \arg \max_{z_j} \vec{v}(z_j) \bar{v}(z_j), \quad (12)$$

we perform a local expectation maximization procedure through message passing [17], [22]. This constraint is enforced for the variables α_l in the Gamma mixture node [23].

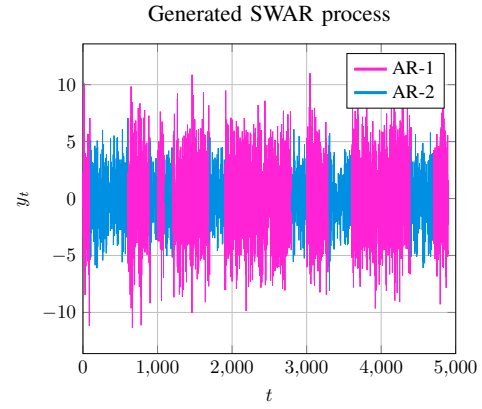


Fig. 2. Example of a generated SwAR signal. The constituent AR processes have been denoted by different colors.

C. Inference in the switching autoregressive model

Inference in the SwAR model of (1)-(8) is performed through a hybrid message passing scheme that includes both sum-product and variational messages. By enforcing different variational constraints on the approximate posterior distributions of the variables in the model, we can obtain different local inference procedures [17]. The graph in Figure 1 submits to a combination of sum-product message passing, (structured) VMP and expectation maximization. Around all deterministic nodes sum-product message passing is performed. Expectation maximization is performed on the edges corresponding to the variables α_l and all other variables submit to (structured) VMP. The message passing update rules for all nodes have already been derived in previous works. Update rules corresponding to the mixture nodes of (4) can be found in [12, Table A2] for the Gaussian mixture node and [13, Table I] for the Gamma mixture node. [12, Table A5] summarizes the update rules for the nodes corresponding to the switching state transition of (5). The update rules corresponding to the Gaussian factor in (1) are summarized in [12, Table A1].

V. EXPERIMENTS

All experiments¹ have been implemented in the Julia programming language [24]. We used the following computer configuration: *Operating system*: macOS Big Sur, *Processor*: 2,7 GHz Quad-Core Intel Core i7, *RAM*: 16GB.

A. Verification experiments

To verify the proposed inference method, we synthesized data from 100 SwAR generative models with the likelihood in (1) with AR order $M = 2$ and $L = 2$ switching states. To ensure the stationarity of the generated processes we resample unstable process configurations. An example of a generated SwAR signal is shown in Figure 4. We used uninformative priors for the transition matrix T and initial switching state c_0 . As for the rest of the model parameters, we used informative priors, i.e., the means of the prior distributions are centered at

¹All experiments are available at <https://github.com/biaslab/swar>.

an ϵ -area ($\epsilon > 0$, $\epsilon^2 \approx 0$) of the means of the corresponding generative distributions. We motivate the usage of informative priors by the non-convexity of the mean-field assumption of our approximate posterior distribution around mixture nodes. This induces multiple solutions for our inference task [25, Ch. 5]. Following the problem definition task in Section III, we seek to obtain the quantities $q(\boldsymbol{\theta}_k|\mathbf{y})$, $q(\gamma_k|\mathbf{y})$, $q(c_k|\mathbf{y})$, $q(\mathbf{T}|\mathbf{y})$ and $q(\boldsymbol{\mu}_l|\mathbf{y})$, $q(\boldsymbol{\Sigma}_l|\mathbf{y})$, $q(\alpha_l|\mathbf{y})$, $q(\beta_l|\mathbf{y})$ for every $l = 1, \dots, L$. The notation $q(\cdot|\mathbf{y})$ refers to the marginal distribution after all observations \mathbf{y} .

Additionally, we want to verify the convergence of the proposed methodology by monitoring the VFE. The inference results are presented in Figures 3 and 4. We evaluate the

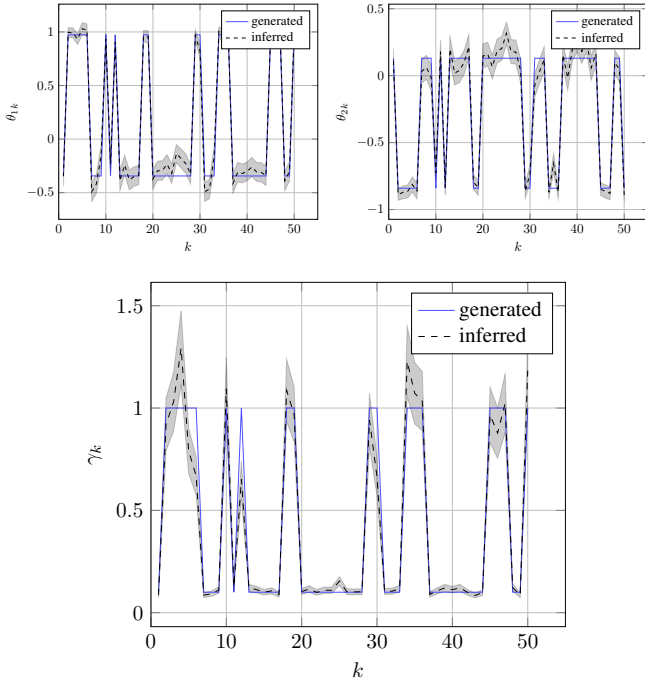


Fig. 3. Inference results on the synthetic dataset. The dashed lines correspond to the expected values of the posterior estimates. The shaded regions correspond to the inferred standard deviation of the approximate posterior distributions around the estimated mean. The solid blue lines correspond to the true underlying values of the latent parameters in the generative processes. (Top) Inference results for the AR coefficients obtained from the joint marginal distribution $q(\boldsymbol{\theta}_k|\mathbf{y})$. (Bottom) Inferred approximate posterior distributions of the precision variables $q(\gamma_k|\mathbf{y})$.

performance of the inference for the switching states procedure by computing a categorical accuracy metric, defined as

$$acc = \frac{tp + tn}{R \cdot K}, \quad (13)$$

where tp , tn are the number of true positive and true negative values, respectively. R corresponds to the number of total synthetic data sets, which in this experiment is set to $R = 100$. In this experiment, we achieved a categorical accuracy of $acc = 0.84$.

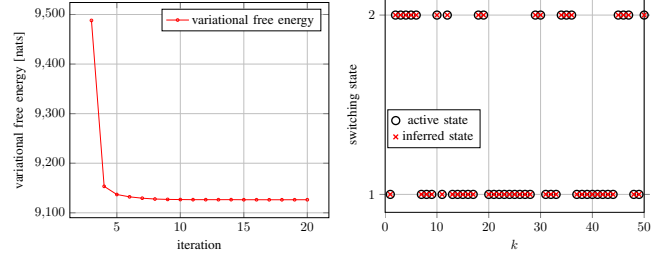


Fig. 4. Inference results on the synthetic dataset. (Left) Evolution of the variational free energy averaged over all generated data sets. (Right) True and inferred evolution of the switching state per frame. Each frame consists of $W = 100$ data points. Circles denote the active switching states that were used to generate the frame. Crosses denote the mode of the inferred switching states.

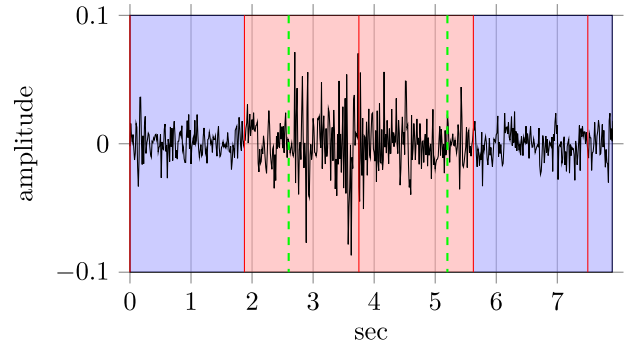


Fig. 5. Inference results for the audio signal. The acoustic signal is represented by a solid black line. Windows of 15000 samples are separated by red solid lines. The green vertical lines correspond to the locations where the underlying acoustic signal changes. The first frame was identified as a train sound (blue region). The two frames in the middle signify a bar sound (red region). The last two frames were classified as a train sound.

B. Validation experiments

To validate the proposed inference procedure, we used 8 seconds of an audio signal, composed of the concatenation of sounds from two different acoustic environments: a train station and a bar. Specifically, we have ≈ 2.6 sec of train sound, followed by ≈ 2.6 sec of bar noise, ending with another train station noise of ≈ 2.6 sec. The sampling frequency was 8 kHz and the audio file is available at <https://github.com/biaslab/swar/data/>. The task is to identify the states of each window, or to classify which acoustic environment is present in the window. In our experiment, we set a maximum window size to 15000 samples (or 1.875 seconds). In this way, our signal breaks into $K = 5$ windows, where the 5th window contains 4000 samples. The choice of 15000 reflects our beliefs about the temporal structure of the signal. In other words, we assume that the switches in the acoustic signal happens at the seconds-level, not at the milliseconds level. We used informative priors for the AR coefficients and precision parameters of the SwAR model. These priors were obtained from performing parameter estimation in the autoregressive model [23]. We have little prior information about the initial state of the audio signal. Thus, we assigned vague (uninformative) priors for

the initial state c_0 and transition matrix T . We present the inference result in Figure 5. Although some frames contain overlapping acoustic signals due to current segmentation, good classification results were achieved through the automated message passing-based inference procedure.

VI. DISCUSSION AND CONCLUSION

We have introduced an SwAR model that includes efficient joint variational tracking of states, parameters, and variational free energy. In this work, we have demonstrated just one way of approximating the posterior distribution of α_l . In particular, we employed a local expectation maximization procedure to estimate the α_l parameter. Although this approach delivers reasonable estimates, it is not suited for online inference scenarios. For these scenarios, one could resort to the moment matching procedure as proposed in [13].

This paper introduced an SwAR model composed from a Gaussian and Gamma mixture model. Owing to the modularity of the factor graph approach, this model can be easily extended and its inference algorithm can be automatically generated based on efforts from previous works. The correctness of the proposed message passing-based inference has been verified on multiple datasets synthesized from the SwAR model. Finally, we demonstrated the convergence of the inference procedure through the minimization of VFE. The proposed model can be easily extended to a latent SwAR model using the update rules of [23], where instead of directly observing y_t we observe a noisy variable $z_t \sim \mathcal{N}(y_t, \gamma)$. We aim at using SwAR model as a module in more complex hierarchical systems in future work.

REFERENCES

- [1] O. Kakusho and M. Yanagida, "Hierarchical AR model for time varying speech signals," in *ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 7, Paris, France, May 1982, pp. 1295–1298.
- [2] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *ICASSP '87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 12, Dallas, TX, USA, Apr. 1987, pp. 177–180.
- [3] J. Gibson, B. Koo, and S. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Transactions on Signal Processing*, vol. 39, no. 8, pp. 1732–1742, Aug. 1991.
- [4] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 373–385, Jul. 1998.
- [5] B. Mesot and D. Barber, "A Bayesian Alternative to Gain Adaptation in Autoregressive Hidden Markov Models," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, vol. 2, Apr. 2007, pp. II-437–II-440, iSSN: 2379-190X.
- [6] D. Barber, *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012. [Online]. Available: <http://www.cs.ucl.ac.uk/staff/d.barber/brml/>
- [7] M. A. Martínez-Beneito, D. Conesa, A. López-Quílez, and A. López-Maside, "Bayesian Markov switching models for the early detection of influenza epidemics," *Statistics in Medicine*, vol. 27, no. 22, pp. 4455–4468, 2008. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3320>
- [8] F. Noman, S.-H. Salleh, C.-M. Ting, S. B. Samdin, H. Ombao, and H. Hussain, "A Markov-Switching Model Approach to Heart Sound Segmentation and Classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 3, pp. 705–716, Mar. 2020.
- [9] P. Ailliot and V. Monbet, "Markov-switching autoregressive models for wind time series," *Environmental Modelling & Software*, vol. 30, pp. 92–101, Apr. 2012. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1364815211002222>
- [10] G. Forney, "Codes on graphs: normal realizations," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 520–548, Feb. 2001. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/910573>
- [11] H.-A. Loeliger, "An introduction to factor graphs," *Signal Processing Magazine, IEEE*, vol. 21, no. 1, pp. 28–41, Jan. 2004. [Online]. Available: <https://ieeexplore.ieee.org/document/1267047>
- [12] T. van de Laar, "Automated Design of Bayesian Signal Processing Algorithms," Ph.D. dissertation, Eindhoven University of Technology, Eindhoven, The Netherlands, 2019.
- [13] A. Podusenko, B. van Erp, D. Bagaev, and Şenöz, İsmail, "Message Passing-Based Inference in the Gamma Mixture Model," in *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*. Gold Coast, Australia: IEEE, Oct. 2021, pp. 1–6. [Online]. Available: <https://ieeexplore.ieee.org/document/9596329/>
- [14] Şenöz, İsmail, A. Podusenko, W. M. Kouw, and B. de Vries, "Bayesian joint state and parameter tracking in autoregressive models," in *Learning for Dynamics and Control*, 2020, p. 9. [Online]. Available: <https://proceedings.mlr.press/v120/senoz20a.html>
- [15] S. Kori, "A factor graph approach to signal modelling, system identification and filtering," Ph.D. dissertation, Swiss Federal Institute of Technology, Zurich, 2005.
- [16] H.-A. Loeliger, J. Dauwels, J. Hu, S. Kori, L. Ping, and F. R. Kschischang, "The Factor Graph Approach to Model-Based Signal Processing," *Proceedings of the IEEE*, vol. 95, no. 6, pp. 1295–1322, Jun. 2007.
- [17] Şenöz, İsmail, T. van de Laar, D. Bagaev, and B. de Vries, "Variational Message Passing and Local Constraint Manipulation in Factor Graphs," *Entropy*, vol. 23, no. 7, p. 807, 2021.
- [18] J. Pearl, "Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach," in *Proceedings of the Second AAAI Conference on Artificial Intelligence*, ser. AAAI'82. Pittsburgh, Pennsylvania: AAAI Press, 1982, pp. 133–136. [Online]. Available: <http://www.aaai.org/Papers/AAAI/1982/AAAI82-032.pdf>
- [19] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on information theory*, vol. 47, no. 2, pp. 498–519, 2001. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=910572
- [20] J. Winn and C. M. Bishop, "Variational message passing," *Journal of Machine Learning Research*, vol. 6, no. 4, pp. 661–694, 2005. [Online]. Available: <http://www.jmlr.org/papers/volume6/winn05a/winn05a.pdf>
- [21] J. Dauwels, "On Variational Message Passing on Factor Graphs," in *IEEE International Symposium on Information Theory*, Nice, France, Jun. 2007, pp. 2546–2550. [Online]. Available: <http://ieeexplore.ieee.org/abstract/document/4557602>
- [22] J. Dauwels, S. Kori, and H.-A. Loeliger, "Expectation maximization as message passing," in *International Symposium on Information Theory, 2005. ISIT 2005. Proceedings*, Sep. 2005, pp. 583–586.
- [23] A. Podusenko, W. M. Kouw, and B. de Vries, "Message Passing-Based Inference for Time-Varying Autoregressive Models," *Entropy*, vol. 23, no. 6, p. 683, Jun. 2021. [Online]. Available: <https://www.mdpi.com/1099-4300/23/6/683>
- [24] J. Bezanson, A. Edelman, S. Karpinski, and V. Shah, "Julia: A Fresh Approach to Numerical Computing," *SIAM Review*, vol. 59, no. 1, pp. 65–98, Jan. 2017. [Online]. Available: <https://epubs.siam.org/doi/abs/10.1137/141000671>
- [25] M. J. Wainwright and M. I. Jordan, "Graphical Models, Exponential Families, and Variational Inference," *Foundations and Trends® in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, Nov. 2008. [Online]. Available: <https://www.nowpublishers.com/article/Details/MAL-001>