Online Bayesian system identification in multivariate autoregressive models via message passing

Tim N. Nisslbeck[1] and Wouter M. Kouw[1]

***Abstract*— We propose a recursive Bayesian estimation procedure for multivariate autoregressive models with exogenous inputs based on message passing in a factor graph. Unlike recursive least-squares, our method produces full posterior distributions for both the autoregressive coefficients and noise precision. The uncertainties regarding these estimates propagate into the uncertainties on predictions for future system outputs, and support online model evidence calculations. We demonstrate convergence empirically on a synthetic autoregressive system and competitive performance on a double mass-spring-damper system.**

I. INTRODUCTION

Autoregressive models capture dynamical systems with simple yet expressive structures [1], [2], [3], [4]. In multivariate autoregressive models with exogenous inputs (MARX), the evolution of the signal incorporates past observations and controls, producing substantial uncertainty during parameter estimation. Bayesian inference procedures can quantify this uncertainty and propagate it towards future predictions [5], [6]. Quantified uncertainty is valuable on its own, but also useful to sensor fusion, optimal experimental design and adaptive control [7], [8], [9], [10], [11]. We present an exact recursive Bayesian estimator whose computation is distributed over a probabilistic graphical model.

Bayesian inference in multivariate autoregressive models has a rich history, especially in econometrics [1], [3]. Scientists typically employ Markov Chain Monte Carlo methods to obtain approximate posterior distributions, but such techniques are too computationally expensive for online system identification or adaptive control systems. Recursive estimators are more suited, but generally lack posterior uncertainty on parameters or produce approximate posterior distributions [2], [5]. We propose an exact recursive Bayesian estimator using the matrix normal Wishart distribution and cast the inference procedure as a message passing algorithm on a factor graph. Factor graphs have several key features. Firstly, they provide a more accessible visual representation of a probabilistic model and the structure of incoming data streams [12], [11]. Secondly, by assigning computation to nodes, model design becomes more modular, allowing inference to be automated and distributed over devices [13], [14]. Thirdly, because messages propagate uncertainty along the graph, the final output uncertainty can be tracked back to individual sources of uncertainty, which allows a prediction to be explained in terms of, for example, volatility versus parameter uncertainty [15]. Lastly, message passing unifies a variety of algorithms, from signal filtering to optimal control and path planning [13], [10], [11].

Our contributions are computational in nature:

- A message passing algorithm is derived for recursive Bayesian inference in MARX models.
- A distribution is derived for predicting future system outputs that incorporates parameter uncertainty.

The performance of our proposed procedure is empirically evaluated in two experimental settings.

*Supported by the Eindhoven Artificial Intelligence Systems Institute.

[1]Nisslbeck and Kouw are with the Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, the Netherlands. Corresponding email: t.n.nisslbeck@tue.nl

II. PROBLEM STATEMENT

We study discrete-time state-space systems that evolve over time according to a state transition function $f : \mathbb{R}^{D_z} \times \mathbb{R}^{D_u} \mapsto \mathbb{R}^{D_z}$ and outputs noisy measurements $y_t \in \mathbb{R}^{D_y}$ through a measurement function $g : \mathbb{R}^{D_z} \mapsto \mathbb{R}^{D_y}$:

$$z_t = f(z_{t\text{-}1}, u_t)\,, \qquad y_t = g(z_t) + e_t\,, \tag{1}$$

with stochastic disturbance $e_t \in \mathbb{R}^{D_y}$ and the goal to predict future outputs y_τ for $\tau > t$ given future inputs u_τ using a model of the system's dynamics.

III. MODEL SPECIFICATION

We consider an order-N MARX model. Let $y_t \in \mathbb{R}^{D_y}$ be a multivariate signal and let

$$\bar{y}_{t\text{-}1} \triangleq \begin{bmatrix} y_{t-1,1} & y_{t-2,1} & \cdots & y_{t-N_y,1} \\ \vdots & \cdots & \cdots & \vdots \\ y_{t-1,D_y} & y_{t-2,D_y} & \cdots & y_{t-N_y,D_y} \end{bmatrix}, \tag{2}$$

be the observation history of memory size N_y. Similarly, we have a matrix of previous values of exogeneous (control) signals

$$\bar{u}_t \triangleq \begin{bmatrix} u_{t,1} & u_{t-1,1} & \cdots & u_{t-N_u+1,1} \\ \vdots & \cdots & \cdots & \vdots \\ u_{t,D_u} & u_{t-1,D_u} & \cdots & u_{t-N_u+1,D_u} \end{bmatrix} \tag{3}$$

with control memory size N_u. We shall reshape both matrices $\bar{y}_{t\text{-}1}$ and $\bar{u}_{t\text{-}1}$ into a $D_x = N_y \times D_y + N_u \times D_u$ vector:

$$x_t \triangleq \left[\mathrm{vec}(\bar{y}_{t\text{-}1})\ \mathrm{vec}(\bar{u}_t)\right]^\intercal . \tag{4}$$

Our likelihood model is

$$p(y_t \mid A, W, x_t) = \mathcal{N}(y_t \mid A^\intercal x_t, W^{-1}) \tag{5}$$

$$= \sqrt{\frac{|W|}{(2\pi)^{D_y}}} \exp\left(-\frac{1}{2}(y_t - A^\intercal x_t)^\intercal W (y_t - A^\intercal x_t)\right). \tag{6}$$

There are two unknowns: a matrix of autoregression coefficients $A \in \mathbb{R}^{D_x \times D_y}$ and a precision matrix $W \in \mathbb{R}^{D_y \times D_y}$. For computational convenience (see Section IV-A), we specify our prior distribution over (A, W) to be a matrix normal Wishart distribution [16, ID: D175]:

$$p(A,W) = \mathcal{MNW}(A, W \mid M_0, \Lambda_0^{-1}, \Omega_0^{-1}, \nu_0) \tag{7}$$

$$= \sqrt{\frac{|\Omega_0|^{D_y}|\Lambda_0|^{\nu_0}}{(2\pi)^{D_x D_y} 2^{\nu_0 D_y}}} \frac{\sqrt{|W|^{\nu_0+D_x-D_y-1}}}{\Gamma_{D_y}(\nu_0/2)} \tag{8}$$

$$\exp\big(-\frac{1}{2}\mathrm{tr}\big[W\big((A-M_0)^\intercal \Lambda_0 (A-M_0)+\Omega_0\big)\big]\big),$$

where $\Gamma_{D_y}(\cdot)$ is the D_y-dimensional multivariate gamma function. The coefficient matrix A follows a matrix normal distribution with mean $M_0 \in \mathbb{R}^{D_x \times D_y}$, row covariance $\Lambda_0^{-1} \in \mathbb{R}^{D_x \times D_x}$, and column covariance W^{-1},

$$p(A \mid W) = \mathcal{MN}(A \,|M_0, \Lambda_0^{-1}, W^{-1}) \tag{9}$$

$$= \sqrt{\frac{|W|^{D_x}|\Lambda_0|^{D_y}}{(2\pi)^{D_x D_y}}} \exp\big(-\frac{1}{2}\mathrm{tr}\big[W(A-M_0)^\intercal \Lambda_0 (A-M_0)\big]\big).$$

The precision matrix W follows a Wishart distribution with scale matrix $\Omega_0^{-1} \in \mathbb{R}^{D_y \times D_y}$ and degrees of freedom $\nu_0 \in \mathbb{R}$

$$p(W) = \mathcal{W}(W \mid \Omega_0^{-1}, \nu_0) \tag{10}$$

$$= \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 D_y}}} \frac{\sqrt{|W|^{\nu_0 - D_y - 1}}}{\Gamma_{D_y}(\frac{\nu_0}{2})} \exp\big(-\frac{1}{2}\mathrm{tr}\big[W\Omega_0\big]\big). \tag{11}$$

Our goal is to infer a posterior distribution over parameters A and W, and later to utilize these parameter posterior distributions to make predictions for future outputs y_τ.

A. Factor graph

The probabilistic graphical model for the recursive model is straightforward, as it constitutes only a prior distribution and a likelihood function. We present a Forney-style factor graph in Fig. 1, where nodes correspond to factors, edges to variables and an edge may only be connected to two nodes [12]. In the graph, time flows from left to right, predictions from top to bottom and corrections from bottom to top. The factor node labeled $\mathcal{MNW}$ represents the matrix normal Wishart prediction distribution with its prior parameters. The dotted box represents the composite likelihood node, consisting of the concatenation operation described in (4), the dot product operation between the matrix of autoregressive coefficients A and the memory x_t, and the stochastic disturbance. The equality node connects the parameters A, W to the likelihood nodes for each t.

IV. INFERENCE

A. Parameter estimation

We are interested in the posterior distribution, which we shall describe recursively:

$$p(A, W \mid \mathcal{D}_t) = \frac{p(y_t \mid A, W, x_t)}{p(y_t \mid \mathcal{D}_t)}\; p(A, W \mid \mathcal{D}_{t\text{-}1}), \tag{12}$$

where the evidence term in the denominator is:

$$p(y_t|u_t, \mathcal{D}_{t\text{-}1}) = \int p(y_t|A,W,x_t) p(A,W|\mathcal{D}_{t\text{-}1}) \mathrm{d}(A,W). \tag{13}$$

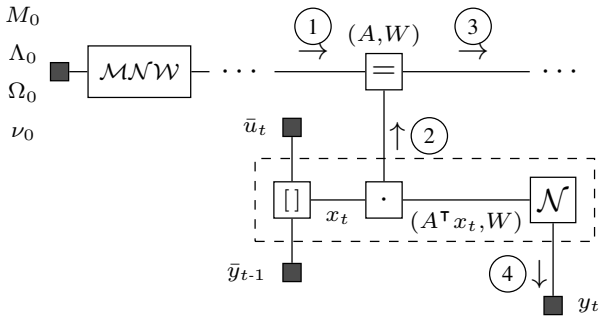

Fig. 1. Forney-style factor graph of the MARX model in recursive form. A matrix normal Wishart node sends a prior message 1 towards an equality node. A likelihood-based message 2 passes upwards from the MARX likelihood node (dotted box), attached to observed variables y_t, $\bar{y}_{t-1}$ and $\bar{u}_t$. Combining the prior-based and likelihood-based message at the equality node yields the posterior (message 3). Message 4 is the posterior predictive distribution for system output.

Lemma 1: The MARX likelihood (5) combined with a matrix normal Wishart prior distribution over MARX coefficient matrix A and precision matrix W (7) yields a matrix normal Wishart distribution:

$$p(A, W \mid \mathcal{D}_t) = \mathcal{MNW}(A, W \mid M_t, \Lambda_t^{-1}, \Omega_t^{-1}, \nu_t) \tag{14}$$

with parameter update rules:

$$\nu_t = \nu_{t\text{-}1} + 1, \tag{15}$$

$$\Lambda_t = \Lambda_{t\text{-}1} + x_t x_t^\intercal, \tag{16}$$

$$M_t = (\Lambda_{t\text{-}1} + x_t x_t^\intercal)^{-1}(\Lambda_{t\text{-}1} M_{t\text{-}1} + x_t y_t^\intercal), \text{ and} \tag{17}$$

$$\Omega_t = \Omega_{t\text{-}1} + y_t y_t^\intercal + M_{t\text{-}1}^\intercal \Lambda_{t\text{-}1} M_{t\text{-}1} \tag{18}$$

$$-(\Lambda_{t\text{-}1} M_{t\text{-}1} + x_t y_t^\intercal)^\intercal (\Lambda_{t\text{-}1} + x_t x_t^\intercal)^{\text{-}1} (\Lambda_{t\text{-}1} M_{t\text{-}1} + x_t y_t^\intercal).$$

See the corresponding proof in Appendix I. This solution can be expressed as a message passing procedure on a factor graph, enabling distributed computation [12].

The circled numbers in Fig. 1 denote messages passed between factor nodes along edges. Message (1) represents the previous posterior belief over coefficients and precision:

$$\overrightarrow{(1)} = p(A, W \mid \mathcal{D}_{t\text{-}1}) \tag{19}$$

$$= \mathcal{MNW}(A, W \mid M_{t\text{-}1}, \Lambda_{t\text{-}1}^{-1}, \Omega_{t\text{-}1}^{-1}, \nu_{t\text{-}1}). \tag{20}$$

The sum-product message from the composite MARX likelihood towards its parameters is the likelihood function itself, re-expressible as a probability distribution over A, W.

Lemma 2: The message from the composite MARX likelihood (5) towards its parameters is matrix normal Wishart distributed:

$$\uparrow (2) = p(y_t \mid A, W, x_t) \tag{21}$$

$$\propto \mathcal{MNW}(A, W \mid \bar{M}_t, \bar{\Lambda}_t^{-1}, \bar{\Omega}_t^{-1}, \bar{\nu}_t). \tag{22}$$

Its parameters are:

$$\bar{\nu}_t = 2 - D_x + D_y, \qquad \bar{\Lambda}_t = x_t x_t^\intercal, \tag{23}$$

$$\bar{M}_t = (x_t x_t^\intercal)^{-1} x_t y_t^\intercal, \qquad \bar{\Omega}_t = 0_{D_y \times D_y}. \tag{24}$$

The proof is in Appendix II.

Message ③ is the result of the multiplication of messages ① and ② in the equality node [12].

Lemma 3: Let p_1, p_2 be two matrix normal Wishart distributions over the same random variables A, W:

$$p_1(A, W) = \mathcal{MNW}(A, W \mid M_1, \Lambda_1^{-1}, \Omega_1^{-1}, \nu_1) \quad (25)$$

$$p_2(A, W) = \mathcal{MNW}(A, W \mid M_2, \Lambda_2^{-1}, \Omega_2^{-1}, \nu_2)\,. \quad (26)$$

Their product is proportional to another matrix normal Wishart distribution:

$$p_1(A,W)p_2(A,W) \propto \mathcal{MNW}(A,W|M_3, \Lambda_3^{-1}, \Omega_3^{-1}, \nu_3) \quad (27)$$

whose parameters are combinations of p_1, p_2's parameters,

$$\nu_3 = \nu_1 + \nu_2 + D_x - D_y - 1\,, \quad (28)$$

$$\Lambda_3 = \Lambda_1 + \Lambda_2\,, \quad (29)$$

$$M_3 = (\Lambda_1 + \Lambda_2)^{-1}(\Lambda_1 M_1 + \Lambda_2 M_2)\,, \quad (30)$$

$$\Omega_3 = \Omega_1 + \Omega_2 + M_1^\intercal \Lambda_1 M_1 + M_2^\intercal \Lambda_2 M_2 \quad (31)$$
$$-(\Lambda_1 M_1 + \Lambda_2 M_2)^\intercal (\Lambda_1 + \Lambda_2)^{-1} (\Lambda_1 M_1 + \Lambda_2 M_2).$$

See Appendix III for the proof.

Theorem 1: The equality node's outgoing message is proportional to the exact recursive posterior distribution;

$$\overrightarrow{③} = \overrightarrow{①} \cdot ② \uparrow \propto \mathcal{MNW}(A, W \mid M_t, \Lambda_t^{-1}, \Omega_t^{-1}, \nu_t) \quad (32)$$

Proof: Combining parameters from the messages in (19) and (21) according the product operation in Lemma 3, yields:

$$\nu_t = \nu_{t\text{-}1} + 1\,, \quad (33)$$

$$\Lambda_t = \Lambda_{t\text{-}1} + x_t x_t^\intercal\,, \quad (34)$$

$$M_t = (\Lambda_{t\text{-}1} + x_t x_t^\intercal)^{-1}(\Lambda_{t\text{-}1} M_{t\text{-}1} + x_t y_t^\intercal)\,, \quad (35)$$

$$\Omega_t = \Omega_{t\text{-}1} + M_{t\text{-}1}^\intercal \Lambda_{t\text{-}1} M_{t\text{-}1} + y_t y_t^\intercal \quad (36)$$
$$-(\Lambda_{t\text{-}1} M_{t\text{-}1} + x_t y_t^\intercal)^\intercal (\Lambda_{t\text{-}1} + x_t x_t^\intercal)^{-1} (\Lambda_{t\text{-}1} M_{t\text{-}1} + x_t y_t^\intercal)$$

These match the parameter update rules of Lemma 1. ∎

B. Output prediction

Predicting future system outputs in our probabilistic model constitutes deriving the posterior predictive distribution, i.e., the marginal distribution of y_τ for $\tau > t$:

$$\downarrow ④ = p(y_\tau \mid u_\tau, \mathcal{D}_t) \quad (37)$$

$$= \int p(y_\tau \mid A, W, x_\tau) p(A, W \mid \mathcal{D}_t)\, \mathrm{d}(A, W)\,. \quad (38)$$

We can utilize the factorisation structure of the parameter posterior to split this into a marginalization over A,

$$p(y_\tau|W, u_\tau, \mathcal{D}_t) = \int p(y_\tau|A, W, x_\tau) p(A|W, \mathcal{D}_t) \mathrm{d}A\,, \quad (39)$$

and a marginalization over W,

$$p(y_\tau \mid u_\tau, \mathcal{D}_t) = \int p(y_\tau \mid W, u_\tau, \mathcal{D}_t) p(W \mid \mathcal{D}_t) \mathrm{d}W\,. \quad (40)$$

Theorem 2: The marginalization of the composite MARX likelihood function (5) over a matrix normal distribution (9) yields a multivariate normal distribution:

$$\int \mathcal{N}\big(y_\tau \mid A^\intercal x_\tau, W^{-1}\big) \mathcal{MN}\big(A \mid M_t, \Lambda_t^{-1}, W^{-1}\big) \mathrm{d}A$$
$$= \mathcal{N}\big(y_\tau \mid M_t^\intercal x_\tau, (\lambda_\tau W)^{-1}\big)\,, \quad (41)$$

where $\lambda_\tau \triangleq (1 + x_\tau^\intercal \Lambda_t^{-1} x_\tau)^{-1}$.

The proof is in Appendix IV. The marginalization of a multivariate normal distribution over a Wishart distribution for its precision parameter yields a multivariate location-scale T-distribution [16, ID: D148]:

$$\int \mathcal{N}\big(y_\tau \mid M_t^\intercal x_\tau, (\lambda_\tau W)^{-1}\big) \mathcal{W}(W \mid \Omega_t^{-1}, \nu_t) \mathrm{d}W$$
$$= \mathcal{T}(y_\tau \mid \mu_\tau, \Psi_\tau^{-1}, \eta_\tau)\,, \quad (42)$$

where $\mu_\tau \triangleq M_t^\intercal x_\tau$, $\eta_\tau \triangleq \nu_t - D_y + 1$, and $\Psi_\tau \triangleq \eta_\tau \Omega_t^{-1} \lambda_\tau$. The posterior predictive distribution provides a recursive uncertainty estimate of the output predictions, which is valuable for decision-making and adaptive control.

V. EXPERIMENTS

We evaluate the MARX estimator on two systems: a multivariate autoregressive system (verification) and a double mass-spring-damper system (validation)[1].

A. Baseline estimator

In both experiments, we compare against a recursive least-squares (RLS) estimator [2]. The coefficient point estimate $\hat{A}_t$ and (initial) inverse sample covariance matrix $P_0 = I_{D_x}$ are updated at each timestep via

$$P_t = P_{t\text{-}1} - P_{t\text{-}1} x_t (1 + x_t^\intercal P_{t\text{-}1} x_t)^{-1} x_t^\intercal P_{t\text{-}1} \quad (43)$$

$$\hat{A}_t = \hat{A}_{t\text{-}1} + P_{t\text{-}1} x_t (1 + x_t^\intercal P_{t\text{-}1} x_t)^{-1} (y_t - \hat{A}_{t\text{-}1}^\intercal x_t)^\intercal\,. \quad (44)$$

This corresponds to a forgetting factor of 1.0, meaning all data points are weighted equally. Predictions are given by $y_\tau = \hat{A}_t^\intercal x_\tau$.

B. Verification

The verification system uses $z_t = x_t$ with $N_y = 2$, $N_u = 3$, and $D_y = D_u = 2$. True coefficients $\tilde{A}$ are generated using Butterworth filters (20 Hz cutoff) for self-connections and Gaussian noise (mean 0, std 0.1) for cross-connections. Disturbances follow $e_t \sim \mathcal{N}(0, \tilde{W}^{-1})$ with $\tilde{W} = \begin{bmatrix} 300 & 100 \\ 100 & 200 \end{bmatrix}$. Training sizes $T_{\text{train}} \in \{2, 4, 6, \ldots, 64\}$ are evaluated over $N_{\text{MC}} = 600$ Monte Carlo runs and $T_{\text{test}} = 100$ test steps. We compare an uninformative prior (MARX-UI, $\Lambda_0 = 1e-4 \cdot I_{D_x}$, $\Omega_0 = 1e-5 \cdot I_{D_y}$) and weakly informative prior (MARX-WI, $\Lambda_0 = 1e-1 \cdot I_{D_x}$, $\Omega_0 = 1e-1 \cdot I_{D_x}$); both with $M_0 = 0_{D_x \times D_y}$, $\nu_0 = D_y + 2$.

At $T_{train} = 2^6$, Fig. 2 shows that MARX-WI consistently achieves better A estimates than MARX-UI and RLS; MARX-UI outperforms RLS after initial instability. Unlike RLS, the MARX estimators also estimate W. Since the MARX estimator is probabilistic, it quantifies uncertainty in its estimations of $\tilde{A}$ and $\tilde{W}$ through precision parameters. The last two subplots in Fig. 2 illustrate the evolution of MARX-WI estimates and their standard deviation ribbons for selected elements of A and W, with uncertainty decreasing over time. All estimators converge to comparable root mean squared errors (RMSE) (with standard errors): $0.284 \pm 5.45e{-}3$ (MARX-WI), $0.289 \pm 5.43e{-}3$ (MARX-UI), and $0.301 \pm 5.50e{-}3$ (RLS).

[1]Code: `https://github.com/biaslab/ECC2025-MARXEFE`

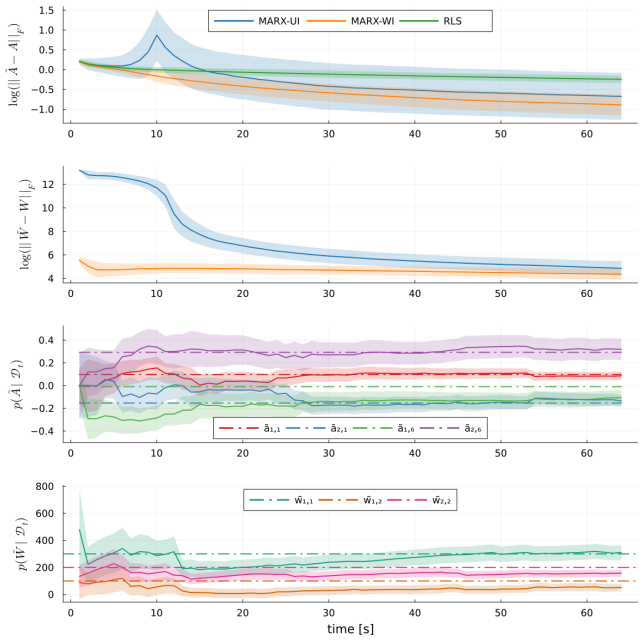

Fig. 2. (**Top**) Log-scale Frobenius-norm differences ($\tilde{A}$ vs. A); (**Second**) $\tilde{W}$ vs W. Time series of selected elements of A (**third**) and W (**bottom**). Shaded areas denote standard errors; horizontal lines indicate true values.

C. Validation

The validation system is a double mass-spring-damper system. Mass m_1 is attached to a fixed base via spring and damper coefficients k_1 and c_1, and mass m_2 is linked to m_1 via k_2 and c_2. The discrete-time dynamics follow a first-order ordinary differential equation (ODE) $I_t \ddot{z}_t = F(z_t, \dot{z}_t, u_t)$ with diagonal inertia matrix I_t with elements m_1 and m_2, generalized coordinates z_t, first and second time derivatives of z_t ($\dot{z}_t$ and $\ddot{z}_t$), and a generalized force function $F(z_t, \dot{z}_t, u_t) = Kz_t + C\dot{z}_t + u_t$ with spring and damping matrices

$$K = \begin{bmatrix} -(k_1+k_2) & k_2 \\ k_2 & -k_2 \end{bmatrix}, \quad C = \begin{bmatrix} -(c_1+c_2) & c_2 \\ c_2 & -c_2 \end{bmatrix}. \quad (45)$$

To evolve the system over discrete time steps $\Delta t = 0.05$, we split its equations of motion into a set of two coupled second-order ODEs $z_\tau = z_t + \Delta t \dot{z}_t$ and $\dot{z}_\tau = \dot{z}_t + \Delta t \ddot{z}_t$, solving the system of ODEs with forward Euler. At $T_{\text{train}} = 2^6$, MARX-WI and MARX-UI achieve lower RMSEs and standard error than RLS: $0.048 \pm 1.26e{-}3$ for (MARX-WI), $0.046 \pm 1.26e{-}3$ (MARX-UI), and $0.074 \pm 2.24e{-}3$ (RLS).

VI. DISCUSSION

The modularity of the factor graph approach offers significant practical benefits. As shown in [12], they enable the visual design of sophisticated algorithms by adding, removing or combining known computational blocks. For example, the factor graph in Fig. 1 can be extended to a time-varying MARX by adding state transition factor nodes between the equality nodes over parameters [13]. In multi-agent robotics, where sensors and actuators are dispersed across platforms, each agent can update its local beliefs via message passing and share only the most informative summaries [17]. This selective communication can minimize bandwidth while rapidly converging to an accurate global model. Recent work underscores the value of transmitting informative variational beliefs in multi-agent settings [18], [19], which allows scalable cooperative learning among heterogeneous agents. Furthermore, distributing computation across nodes opens up promising avenues in federated system identification and multi-robot coordination, especially under privacy or bandwidth constraints [9], [7], [20]. Overall, the factor graph approach fosters robust, scalable, and efficient inference in distributed sensing, control and robotics, as local computations are fused to form a coherent global model.

VII. CONCLUSIONS

We presented a recursive Bayesian estimation procedure for multivariate autoregressive models with exogenous inputs formulated as message passing on a Forney-style factor graph. It produces matrix-variate posterior distributions for the coefficients and noise precision, whose uncertainty is propagated into future system output predictions.

APPENDIX I
PARAMETER ESTIMATION

Proof: The functional form of the likelihood is

$$p(y_t \mid A, W, x_t) \propto \sqrt{|W|} \exp\left(-\frac{1}{2}\text{tr}\left[W L_t\right]\right), \quad (46)$$

where $L_t \triangleq (y_t - A^\intercal x_t)(y_t - A^\intercal x_t)^\intercal$. The prior is

$$p(A,W|\mathcal{D}_{t\text{-}1}) \propto \sqrt{|W|^{\nu_{t\text{-}1}+\bar{D}}} \exp(-\frac{1}{2}\text{tr}[W(H_{t\text{-}1}+\Omega_{t\text{-}1})]), \quad (47)$$

where $H_{t\text{-}1} \triangleq (A - M_{t\text{-}1})^\intercal \Lambda_{t\text{-}1}(A - M_{t\text{-}1})$ and $\bar{D} \triangleq D_x - D_y - 1$. The posterior is proportional to the prior times likelihood:

$$p(A, W \mid \mathcal{D}_t) \propto p(y_t \mid A, W, x_t)\, p(A, W \mid \mathcal{D}_{t\text{-}1}) \quad (48)$$

$$\propto \sqrt{|W|^{\nu_{t\text{-}1}+1+\bar{D}}} \exp\left(-\frac{1}{2}\text{tr}\left[W(L_t + H_{t\text{-}1} + \Omega_{t\text{-}1})\right]\right). \quad (49)$$

We expand the first terms in the exponent and group them:

$$\begin{aligned} L_t + H_{t\text{-}1} &= y_t y_t^\intercal - y_t x_t^\intercal A - A^\intercal x_t y_t^\intercal + A^\intercal x_t x_t^\intercal A \quad (50)\\ &\quad + A^\intercal \Lambda_{t\text{-}1} A - A^\intercal \Lambda_{t\text{-}1} M_{t\text{-}1} - M_{t\text{-}1}^\intercal \Lambda_{t\text{-}1} A + M_{t\text{-}1}^\intercal \Lambda_{t\text{-}1} M_{t\text{-}1} \\ &= A^\intercal(\Lambda_{t\text{-}1} + x_t x_t^\intercal)A - A^\intercal(x_t y_t^\intercal + \Lambda_{t\text{-}1} M_{t\text{-}1}) \quad (51)\\ &\quad - (M_{t\text{-}1}^\intercal \Lambda_{t\text{-}1} + y_t x_t^\intercal)A + y_t y_t^\intercal + M_{t\text{-}1}^\intercal \Lambda_{t\text{-}1} M_{t\text{-}1}. \end{aligned}$$

Let $\Lambda_t \triangleq \Lambda_{t\text{-}1} + x_t x_t^\intercal$, $\xi_t \triangleq x_t y_t^\intercal + \Lambda_{t\text{-}1} M_{t\text{-}1}$ and $M_t \triangleq \Lambda_t^{-1}\xi_t$. Adding and subtracting $\xi_t^\intercal \Lambda_t^{-1} \xi_t$ to (51) yields:

$$\begin{aligned} L_t + H_{t\text{-}1} &= A^\intercal \Lambda_t A - A^\intercal \xi_t - \xi_t^\intercal A + \xi_t^\intercal \Lambda_t^{-1} \xi_t \quad (52)\\ &\quad - \xi_t^\intercal \Lambda_t^{-1} \xi_t + y_t y_t^\intercal + M_{t\text{-}1}^\intercal \Lambda_{t\text{-}1} M_{t\text{-}1} \\ &= (A - \Lambda_t^{-1}\xi_t)^\intercal \Lambda_t (A - \Lambda_t^{-1}\xi_t) \quad (53)\\ &\quad - M_t^\intercal \Lambda_t M_t + y_t y_t^\intercal + M_{t\text{-}1}^\intercal \Lambda_{t\text{-}1} M_{t\text{-}1}. \end{aligned}$$

Plugging the above into (49), we recognize the functional form of the matrix normal Wishart distribution:

$$\begin{aligned} &\sqrt{|W|^{\nu_t+\bar{D}}} \exp\left(-\frac{1}{2}\text{tr}\left[W((A-M_t)^\intercal \Lambda_t (A-M_t) + \Omega_t)\right]\right) \\ &\quad \propto \mathcal{MNW}(A, W \mid M_t, \Lambda_t^{-1}, \Omega_t^{-1}, \nu_t) \quad (54) \end{aligned}$$

whose parameters are

$$\nu_t = \nu_{t\text{-}1} + 1\,, \tag{55}$$
$$\Lambda_t = \Lambda_{t\text{-}1} + x_t x_t^\intercal\,, \tag{56}$$
$$M_t = (\Lambda_{t\text{-}1} + x_t x_t^\intercal)^{-1}(\Lambda_{t\text{-}1} M_{t\text{-}1} + x_t y_t^\intercal)\,, \text{ and} \tag{57}$$
$$\Omega_t = \Omega_{t\text{-}1} + y_t y_t^\intercal + M_{t\text{-}1}^\intercal \Lambda_{t\text{-}1} M_{t\text{-}1} - M_t^\intercal \Lambda_t M_t\,. \tag{58}$$

This concludes the proof. ∎

Appendix II
Backwards message from likelihood

Proof: The MARX likelihood function is:

$$p(y_t \mid A, W, x_t) \propto \sqrt{|W|}\exp\big(-\frac{1}{2}\mathrm{tr}\big[W L_t\big]\big)\,, \tag{59}$$

where the completed square is

$$L_t \triangleq (y_t - A^\intercal x_t)(y_t - A^\intercal x_t)^\intercal \tag{60}$$
$$= y_t y_t^\intercal - A^\intercal x_t y_t^\intercal - y_t x_t^\intercal A + A^\intercal x_t x_t^\intercal A\,. \tag{61}$$

Let $\bar{\Lambda}_t \triangleq x_t x_t^\intercal$, $\bar{\xi}_t \triangleq x_t y_t^\intercal$ and $\bar{M}_t = \bar{\Lambda}_t^{-1}\bar{\xi}_t$. Then adding and subtracting $\bar{\xi}_t^\intercal \bar{\Lambda}_t \bar{\xi}_t$ allows us to rewrite the square in terms of A:

$$L_t + \bar{\xi}_t^\intercal \bar{\Lambda}_t^{-1}\bar{\xi}_t - \bar{\xi}_t^\intercal \bar{\Lambda}_t^{-1}\bar{\xi}_t \tag{62}$$
$$= y_t y_t^\intercal + (A - \bar{M}_t)^\intercal \bar{\Lambda}_t (A - \bar{M}_t) - \bar{\xi}_t^\intercal \bar{\Lambda}_t^{-1}\bar{\xi}_t\,.$$

The two remaining terms cancel:

$$y_t y_t^\intercal - \bar{\xi}_t^\intercal \bar{\Lambda}_t^{-1}\bar{\xi}_t = y_t y_t^\intercal - y_t x_t^\intercal (x_t x_t^\intercal)^{-1} x_t y_t^\intercal \tag{63}$$
$$= y_t y_t^\intercal - y_t I y_t^\intercal = 0_{D_y \times D_y}\,. \tag{64}$$

If we define $\bar{\nu}_t \triangleq 1 - \bar{D}$ for $\bar{D} = D_x + D_y + 1$ and $\bar{\Omega}_t \triangleq 0_{D_y \times D_y}$, then we may recognize the functional form of a matrix normal Wishart in (59);

$$p(y_t|A, W, x_t) \tag{65}$$
$$\propto \sqrt{|W|^{\bar{\nu}_t + \bar{D}}}\exp\big(-\frac{1}{2}\mathrm{tr}\big[W((A-\bar{M}_t)^\intercal \bar{\Lambda}_t (A-\bar{M}_t) + \bar{\Omega}_t)\big]\big)$$
$$\propto \mathcal{MNW}(A, W \mid \bar{M}_t, \bar{\Lambda}_t^{-1}, \bar{\Omega}_t^{-1}, \bar{\nu}_t)\,. \tag{66}$$

This concludes the proof. ∎

Appendix III
Product of matrix normal Wishart distributions

Proof: Let p_1, p_2 be two matrix normal Wishart distributions over the same random variables A, W:

$$p_1(A, W) = \mathcal{MNW}(A, W \mid M_1, \Lambda_1^{-1}, \Omega_1^{-1}, \nu_1) \tag{67}$$
$$p_2(A, W) = \mathcal{MNW}(A, W \mid M_2, \Lambda_2^{-1}, \Omega_2^{-1}, \nu_2)\,. \tag{68}$$

Their product is proportional to:

$$p_1(A,W)\, p_2(A, W)$$
$$\propto \sqrt{|W|^{\nu_1 + \bar{D}}}\exp\big(-\frac{1}{2}\mathrm{tr}\big[W L_1\big]\big) \tag{69}$$
$$\sqrt{|W|^{\nu_2 + \bar{D}}}\exp\big(-\frac{1}{2}\mathrm{tr}\big[W L_2\big]\big)$$
$$= \sqrt{|W|^{\nu_3 + \bar{D}}}\exp\big(-\frac{1}{2}\mathrm{tr}\big[W(L_1 + L_2)\big]\big) \tag{70}$$

for $\bar{D} \triangleq D_x - D_y - 1$, $\nu_3 \triangleq \nu_1 + \nu_2 + D_x - D_y - 1$ and $L_i \triangleq (A - M_i)^\intercal \Lambda_i (A - M_i) + \Omega_i$. The sum of L_i is:

$$L_1 + L_2 = A^\intercal(\Lambda_1 + \Lambda_2)A \tag{71}$$
$$- A^\intercal(\Lambda_1 M_1 + \Lambda_2 M_2) - (M_1^\intercal \Lambda_1 + M_2^\intercal \Lambda_2)A$$
$$+ M_1^\intercal \Lambda_1 M_1 + M_2^\intercal \Lambda_2 M_2 + \Omega_1 + \Omega_2\,.$$

Let $\Lambda_3 \triangleq \Lambda_1 + \Lambda_2$ and $\xi_3 \triangleq \Lambda_1 M_1 + \Lambda_2 M_2$. Then:

$$(A - \Lambda_3^{-1}\xi_3)^\intercal \Lambda_3 (A - \Lambda_3^{-1}\xi_3) =$$
$$A^\intercal \Lambda_3 A - A^\intercal \xi_3 - \xi_3^\intercal A + \xi_3^\intercal \Lambda_3^{-1}\xi_3\,. \tag{72}$$

Using $M_3 \triangleq \Lambda_3^{-1}\xi_3$, (71) can be written as:

$$p_1(A, W) p_2(A, W) \tag{73}$$
$$\propto \sqrt{|W|^{\nu_3 + \bar{D}}}\exp\big(-\frac{1}{2}\mathrm{tr}\big[W\big((A - M_3)^\intercal \Lambda_3 (A - M_3)$$
$$- \xi_3^\intercal \Lambda_3^{-1}\xi_3 + M_1^\intercal \Lambda_1 M_1 + M_2^\intercal \Lambda_2 M_2 + \Omega_1 + \Omega_2\big)\big]\big)$$

Note that $\xi_3^\intercal \Lambda_3^{-1}\xi_3 = \xi_3^\intercal \Lambda_3^{-1}\Lambda_3\Lambda_3^{-1}\xi_3 = M_3^\intercal \Lambda_3 M_3$. Let

$$\Omega_3 \triangleq \Omega_1 + \Omega_2 + M_1^\intercal \Lambda_1 M_1 + M_2^\intercal \Lambda_2 M_2 - M_3^\intercal \Lambda_3 M_3\,. \tag{74}$$

Then (73) may be recognized as an unnormalized matrix normal Wishart:

$$\sqrt{|W|^{\nu_3 + \bar{D}}}\exp\big(-\frac{1}{2}\mathrm{tr}\big[W\big((A - M_3)^\intercal \Lambda_3 (A - M_3) + \Omega_3\big)\big]\big)$$
$$\propto \mathcal{MNW}\big(A, W \mid M_3, \Lambda_3^{-1}, \Omega_3^{-1}, \nu_3\big)\,. \tag{75}$$

As such, the product of two matrix normal Wishart distributions is proportional to another matrix normal Wishart distribution. ∎

Appendix IV
Marginalization over matrix normal dist.

Proof: Let $L_\cdot \triangleq (y_\cdot - A^\intercal x_\cdot)(y_\cdot - A^\intercal x_\cdot)^\intercal$ and $H_\cdot \triangleq (A - M_\cdot)^\intercal \Lambda_\cdot (A - M_\cdot)$ where the subscript $\cdot$ indicates a time index. The marginalization over A is:

$$p(y_\tau|u_\tau, W; \mathcal{D}_t) = \int p(y_\tau|A, W, x_\tau) p(A|W, \mathcal{D}_t)\,\mathrm{d}A \tag{76}$$
$$= \int \mathcal{N}\big(y_\tau|A^\intercal x_\tau, W^{-1}\big)\mathcal{MN}\big(A|M_t, \Lambda_t^{-1}, W^{-1}\big)\mathrm{d}A \tag{77}$$
$$= \sqrt{(2\pi)^{-D_y(D_x+1)}|W|^{D_x+1}|\Lambda_t|^{D_y}}$$
$$\int \exp\big(-\frac{1}{2}\mathrm{tr}\big[W(L_\tau + H_t)\big]\big)\mathrm{d}A. \tag{78}$$

Expanding L_τ and H_t and adding them yields:

$$L_\tau + H_t = y_\tau y_\tau^\intercal + M_t^\intercal \Lambda_t M_t + A^\intercal(\Lambda_t + x_\tau x_\tau^\intercal)A \tag{79}$$
$$- A^\intercal(\Lambda_t M_t + x_\tau y_\tau^\intercal) - (\Lambda_t M_t + x_\tau y_\tau^\intercal)^\intercal A.$$

Let $\Lambda_\tau \triangleq \Lambda_t + x_\tau x_\tau^\intercal$, $\xi_\tau \triangleq \Lambda_t M_t + x_\tau y_\tau^\intercal$ and $M_\tau \triangleq \Lambda_\tau^{-1}\xi_\tau$. Completing the square gives:

$$L_\tau + H_t = H_\tau - M_\tau^\intercal \Lambda_\tau M_\tau + y_\tau y_\tau^\intercal + M_t^\intercal \Lambda_t M_t\,. \tag{80}$$

Plugging this result into the integral in (78) gives:

$$\int \exp\big(-\frac{1}{2}\mathrm{tr}\big[W(L_\tau + H_t)\big]\big)\mathrm{d}A = \int \exp\big(-\frac{1}{2}\mathrm{tr}\big[W H_\tau\big]\big)\mathrm{d}A$$
$$\cdot \exp\big(-\frac{1}{2}\mathrm{tr}\big[W(y_\tau y_\tau^\intercal + M_t^\intercal \Lambda_t M_t - M_\tau^\intercal \Lambda_\tau M_\tau)\big]\big). \tag{81}$$

We can recognize the integrand as the functional form of a matrix normal distribution. Thus, the integral evaluates to its inverse normalization factor:

$$\int \exp\big(-\frac{1}{2}\mathrm{tr}\big[WH_\tau\big]\big)\, dA = \sqrt{\frac{(2\pi)^{D_y D_x}}{|W|^{D_x}|\Lambda_\tau|^{D_y}}}\,. \tag{82}$$

Using this result, the marginalization over A becomes:

$$\begin{aligned} p(y_\tau \mid u_\tau, W; \mathcal{D}_t) &= \sqrt{(2\pi)^{-D_y}|\Lambda_t|^{D_y}|\Lambda_\tau|^{-D_y}|W|} \\ &\cdot \exp\big(-\frac{1}{2}\mathrm{tr}\big[W(y_\tau y_\tau^\intercal + M_t^\intercal \Lambda_t M_t - M_\tau^\intercal \Lambda_\tau M_\tau)\big]\big). \end{aligned} \tag{83}$$

Note that, under the matrix determinant lemma,

$$|\Lambda_\tau| = |\Lambda_t + x_\tau x_\tau^\intercal| = |\Lambda_t|(1 + x_\tau^\intercal \Lambda_t^{-1} x_\tau)\,, \tag{84}$$

which implies that the product of determinants is

$$|\Lambda_t|^{D_y}|\Lambda_\tau|^{-D_y} = \big(1 + x_\tau^\intercal \Lambda_t^{-1} x_\tau\big)^{-D_y}\,. \tag{85}$$

Let $\lambda_\tau \triangleq (1 + x_\tau^\intercal \Lambda_t^{-1} x_\tau)^{-1}$. As W is D_y-dimensional, $|W|\lambda_\tau^{D_y} = |W\lambda_\tau|$. Furthermore, note that

$$\begin{aligned} M_\tau^\intercal \Lambda_\tau M_\tau = {} & M_t^\intercal \Lambda_t (x_\tau x_\tau^\intercal + \Lambda_t)^{-1} \Lambda_t M_t \\ & + y_\tau x_\tau^\intercal (x_\tau x_\tau^\intercal + \Lambda_t)^{-1} \Lambda_t M_t \\ & + M_t^\intercal \Lambda_t (x_\tau x_\tau^\intercal + \Lambda_t)^{-1} x_\tau y_\tau^\intercal \\ & + y_\tau x_\tau^\intercal (x_\tau x_\tau^\intercal + \Lambda_t)^{-1} x_\tau y_\tau^\intercal. \end{aligned} \tag{86}$$

Combining this with the other terms in the trace gives:

$$\begin{aligned} & y_\tau y_\tau^\intercal + M_t^\intercal \Lambda_t M_t - M_\tau^\intercal \Lambda_\tau M_\tau \\ & \quad = M_t^\intercal \Lambda_t \big(I - (x_\tau x_\tau^\intercal + \Lambda_t)^{-1} \Lambda_t\big) M_t \\ & \qquad - y_\tau x_\tau^\intercal (x_\tau x_\tau^\intercal + \Lambda_t)^{-1} \Lambda_t M_t \\ & \qquad - M_t^\intercal \Lambda_t (x_\tau x_\tau^\intercal + \Lambda_t)^{-1} x_\tau y_\tau^\intercal \\ & \qquad + y_\tau \big(1 - x_\tau^\intercal (x_\tau x_\tau^\intercal + \Lambda_t)^{-1} x_\tau\big) y_\tau^\intercal\,. \end{aligned} \tag{87}$$

Applying the Sherman-Morrison formula and re-arranging terms yields the following simplifications:

$$\big(1 - x_\tau^\intercal (x_\tau x_\tau^\intercal + \Lambda_t)^{-1} x_\tau\big) = \lambda_\tau \tag{88}$$

$$I - (x_\tau x_\tau^\intercal + \Lambda_t)^{-1}\Lambda_t = \lambda_\tau \Lambda_t^{-1} x_\tau x_\tau^\intercal \Lambda_t^{-1} \tag{89}$$

$$\Lambda_t (x_\tau x_\tau^\intercal + \Lambda_t)^{-1} x_\tau = x_\tau \lambda_\tau\,. \tag{90}$$

Using these three simplifications, we have:

$$\begin{aligned} & \mathrm{tr}\Big[W\big(y_\tau y_\tau^\intercal + M_t^\intercal \Lambda_t M_t - M_\tau^\intercal \Lambda_\tau M_\tau\big)\Big] \\ & \qquad = \big(y_\tau - M_t^\intercal x_\tau\big)^\intercal W \lambda_\tau \big(y_\tau - M_t^\intercal x_\tau\big). \end{aligned} \tag{91}$$

Plugging (91) into (78) yields

$$\begin{aligned} & p(y_\tau \mid u_\tau, W; \mathcal{D}_t) \\ & = \sqrt{\frac{|W\lambda_\tau|}{(2\pi)^{D_y}}} \exp\big(-\frac{\lambda_\tau}{2}(y_\tau - M_t^\intercal x_\tau)^\intercal W (y_\tau - M_t^\intercal x_\tau)\big) \\ & = \mathcal{N}\big(y_\tau \mid M_t^\intercal x_\tau, (W\lambda_\tau)^{-1}\big)\,. \end{aligned} \tag{92, 93}$$

This concludes the proof. ∎

References

[1] G. C. Tiao and A. Zellner, "On the Bayesian estimation of multivariate regression," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 26, no. 2, pp. 277–285, 1964.

[2] E. J. Hannan, A. McDougall, and D. S. Poskitt, "Recursive estimation of autoregressions," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 51, no. 2, pp. 217–233, 1989.

[3] S. Karlsson, "Forecasting with Bayesian vector autoregression," *Handbook of Economic Forecasting*, vol. 2, pp. 791–897, 2013.

[4] T. N. Nisslbeck and W. M. Kouw, "Coupled autoregressive active inference agents for control of multi-joint dynamical systems," in *International Workshop on Active Inference*, Springer, 2024.

[5] W. Penny and L. Harrison, "Multivariate autoregressive models," *Statistical Parametric Mapping: The Analysis of Functional Brain Images*, pp. 534–540, 2007.

[6] S. M. Shaarawy and S. S. Ali, "Bayesian identification of multivariate autoregressive processes," *Communications in Statistics—Theory and Methods*, vol. 37, no. 5, pp. 791–802, 2008.

[7] F. Castaldo and F. A. Palmieri, "A multi-camera multi-target tracker based on factor graphs," in *IEEE International Symposium on Innovations in Intelligent Systems and Applications*, pp. 131–137, IEEE, 2014.

[8] K. Chaloner and I. Verdinelli, "Bayesian experimental design: A review," *Statistical Science*, pp. 273–304, 1995.

[9] D.-N. Ta, M. Kobilarov, and F. Dellaert, "A factor graph approach to estimation and model predictive control on unmanned aerial vehicles," in *International Conference on Unmanned Aircraft Systems*, pp. 181–188, IEEE, 2014.

[10] C. Hoffmann and P. Rostalski, "Linear optimal control on factor graphs—a message passing perspective—," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 6314–6319, 2017.

[11] F. A. Palmieri, K. R. Pattipati, G. Di Gennaro, G. Fioretti, F. Verolla, and A. Buonanno, "A unifying view of estimation and control using belief propagation with application to path planning," *IEEE Access*, vol. 10, pp. 15193–15216, 2022.

[12] H.-A. Loeliger, J. Dauwels, J. Hu, S. Korl, 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, 2007.

[13] 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, 2021.

[14] D. Bagaev and B. de Vries, "Reactive message passing for scalable bayesian inference," *Scientific Programming*, vol. 2023, no. 1, p. 6601690, 2023.

[15] F. Lecue, "On the role of knowledge graphs in explainable AI," *Semantic Web*, vol. 11, no. 1, pp. 41–51, 2020.

[16] J. Soch, C. Allefeld, T. J. Faulkenberry, M. Pavlovic, K. Petrykowski, K. Sarıtaş, S. Balkus, A. Kipnis, H. Atze, and O. A. Martin, "The Book of Statistical Proofs (Version 2023)," Jan. 2024.

[17] M. R. Damgaard, R. Pedersen, and T. Bak, "Study of variational inference for flexible distributed probabilistic robotics," *Robotics*, vol. 11, no. 2, p. 38, 2022.

[18] B. C. Tedeschini, M. Brambilla, and M. Nicoli, "Message passing neural network versus message passing algorithm for cooperative positioning," *IEEE Transactions on Cognitive Communications and Networking*, vol. 9, no. 6, pp. 1666–1676, 2023.

[19] Y. Zhang, W. Xu, A. Liu, and V. Lau, "Message passing based wireless federated learning via analog message aggregation," in *2024 IEEE/CIC International Conference on Communications in China (ICCC)*, pp. 2161–2166, IEEE, 2024.

[20] B. van Erp, D. Bagaev, A. Podusenko, İ. Şenöz, and B. de Vries, "Multi-agent trajectory planning with NUV priors," in *American Control Conference*, pp. 2766–2771, IEEE, 2024.