entropy

MDPI

Article

Factor Graph-Based Online Bayesian Identification and Component Evaluation for Multivariate Autoregressive Exogenous Input Models †

Tim N. Nisslbeck * and Wouter M. Kouw

Department of Electrical Engineering, Eindhoven University of Technology, 5612 AZ Eindhoven, The Netherlands; w.m.kouw@tue.nl

* Correspondence: t.n.nisslbeck@tue.nl

† This paper is an extended version of our paper published in the proceedings of the IEEE European Control Conference, held at Thessaloniki, Greece, 24–27 June 2025.

Abstract

We present a Forney-style factor graph representation for the class of multivariate autoregressive models with exogenous inputs, and we propose an online Bayesian parameter-identification procedure based on message passing within this graph. We derive message-update rules for (1) a custom factor node that represents the multivariate autoregressive likelihood function and (2) the matrix normal Wishart distribution over the parameters. The flow of messages reveals how parameter uncertainty propagates into predictive uncertainty over the system outputs and how individual factor nodes and edges contribute to the overall model evidence. We evaluate the message-passing-based procedure on (i) a simulated autoregressive system, demonstrating convergence, and (ii) on a benchmark task, demonstrating strong predictive performance.

Keywords: Bayesian inference; probabilistic graphical models; message passing; system identification; stochastic systems; autoregressive models

Academic Editors: Martin Trapp and Pierre Alquier

Received: 9 March 2025
Revised: 5 June 2025
Accepted: 9 June 2025
Published: 26 June 2025

Citation: Nisslbeck, T.N.; Kouw, W.M. Factor Graph-Based Online Bayesian Identification and Component Evaluation for Multivariate Autoregressive Exogenous Input Models. *Entropy* **2025**, *27*, 679. https://doi.org/10.3390/e27070679

1. Introduction

Autoregressive models provide a simple yet powerful framework for capturing dynamical systems [1–5]. Multivariate autoregressive models with exogenous inputs (MARX) exhibit a complex dependence structure. Each component of the vector signal evolves as a weighted combination of (i) its own past observations, (ii) other components, and (iii) an exogenous vector-valued input signal [6,7]. This intricate dependence structure generates significant uncertainty in parameter estimation.

Bayesian inference offers a principled approach for quantifying and propagating this uncertainty into predictions for future system outputs [8,9]. Moreover, uncertainty quantification enables the incorporation of information-theoretic quantities into cost functions, which is useful for optimal experimental design and adaptive control [10,11]. Markov Chain Monte Carlo techniques are typically employed to approximate posterior distributions. However their computational cost makes them impractical for large-scale real-time applications such as online system identification and adaptive control. In contrast, exact and variational inference methods provide full posterior distributions over parameters, thereby enabling robust decision-making under uncertainty [12,13]. This capability is particularly crucial in safety-critical applications, such as robotics, where understanding uncertainty is as important as making accurate predictions.

To address this challenge, we introduce an exact recursive Bayesian estimator that maintains a full posterior distribution and is computationally efficient. Recursive estimators offer a scalable alternative to batch estimators, but they either lack posterior uncertainty over parameters or rely on approximations [3,8]. Shaarawy and Ali proposed an exact recursive Bayesian estimator based on the matrix normal Wishart distribution, demonstrating its effectiveness for system identification [9]. We extend their approach by casting the inference procedure as a message-passing algorithm on a factor graph, thereby improving both computational efficiency and interpretability.

Factor graphs are graphical tools that capture the probabilistic relationships between random variables [14]. Many algorithms, including inference, can be formulated as message passing on a factor graph. Thus, message passing on factor graphs provides a structured and scalable framework for Bayesian inference, offering several key advantages over conventional inference frameworks [15–17]. We specifically consider Forney-style factor graphs, for their simplicity and compact visual representation [18]. First, factor graphs offer an intuitive representation of probabilistic models and data flow by depicting distinct probabilistic relationships as separate factor nodes that explicitly capture dependencies between variables [15,17]. This structured representation makes the inference process more interpretable and supports a more flexible model design, contributing to explainable artificial intelligence [19,20]. Second, message passing on factor graphs enables distributed computation by structuring inference into localized update rules at each node [21]. In particular, casting inference as message passing on a factor graph can enable federated learning, which accelerates learning in a multi-agent setting where physically separated agents share likelihood messages for joint parameter estimation [22]. This formulation significantly reduces the computational complexity compared to traditional recursive methods, making real-time Bayesian inference more tractable in large-scale settings [23,24]. Localized updates facilitate the efficient propagation of uncertainty throughout the graph, allowing for the attribution of uncertainty to specific sources, for example, distinguishing between prediction uncertainty arising from the likelihood model versus uncertainty in the inferred parameters. This fine-grained decomposition of uncertainties further enables a novel evaluation of model performance: the negative log-model evidence (surprisal) can be decomposed into contributions from individual nodes and edges in the factor graph. By analyzing how these contributions evolve over time, one gains detailed insights into the learning dynamics during system identification, thus linking model evaluation directly to the underlying probabilistic structure. Lastly, message passing unifies a broad class of algorithms, spanning signal filtering, optimal control, and path planning [14,17,24,25], making it a computationally efficient tool for probabilistic reasoning in large-scale problems. Overall, by leveraging this structured inference technique, our approach not only enhances Bayesian inference for dynamical systems but also yields more interpretable, scalable, and computationally efficient probabilistic machine learning models.

In summary, our key contributions are as follows:

- We derive a message-passing algorithm for exact recursive Bayesian inference in MARX models, maintaining full posterior distributions while ensuring computational efficiency.
- We extend the inference framework to predict future system outputs that explicitly account for parameter uncertainty, improving robustness for real-time applications.
- We introduce a novel model evaluation method by decomposing the negative log-model evidence (surprisal) into contributions from individual nodes and edges in the factor graph, providing insights into uncertainty and learning dynamics.
- We demonstrate the effectiveness of our approach through empirical evaluations on (i) a synthetic MARX system with known parameters for verification, and (ii) two synthetic

dynamical systems with unknown parameters: a double mass-spring-damper system and a nonlinear double pendulum system.

The remainder of this paper is organized as follows. In Section 2, we formally describe the class of the discrete-time dynamical system considered. In Section 3, we present our probabilistic MARX model and its representation using Forney-style factor graphs. In Section 4, we detail the message-passing algorithm for recursive Bayesian inference, including both parameter estimation and predictive inference. In Section 5, we introduce our novel evaluation method based on decomposing surprisal. In Section 6, we demonstrate the effectiveness of our approach on synthetic system identification tasks. In Section 7, we discuss the computational benefits, interpretability, and broader implications of our method. Finally, in Section 8, we conclude this paper.

2. Problem Statement

We consider discrete-time dynamical systems, represented by a state $z_k \in \mathbb{R}^{D_z}$ and driven by a control signal $u_k \in \mathbb{R}^{D_u}$. These systems evolve according to a state transition function $f : \mathbb{R}^{D_z} \times \mathbb{R}^{D_u} \mapsto \mathbb{R}^{D_z}$. At each time step, we observe a noisy measurement $y_k \in \mathbb{R}^{D_y}$ of the state via a measurement function $g : \mathbb{R}^{D_z} \mapsto \mathbb{R}^{D_y}$. This can be expressed as a state–space model of the form:

$$z_k = f(z_{k-1}, u_k)\,, \qquad y_k = g(z_k) + e_k\,,$$

where $e_k \in \mathbb{R}^{D_y}$ is a stochastic disturbance. Our objective is to predict future observations y_t for $t > k$, given future inputs u_t, without prior knowledge about the system dynamics.

3. Model Specification

To address the problem defined in Section 2, we propose a probabilistic model that enables recursive learning and prediction of future observations in a partially observed dynamical system. Specifically, we assume that the unknown system can be approximated by a multivariate autoregressive model with exogenous inputs of order N, denoted as MARX(N). Let $y_k \in \mathbb{R}^{D_y}$ denote the D_y-dimensional observation at time step k. We collect the past N_y outputs into the matrix

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

and, similarly, the most recent N_u control inputs into

$$\bar{u}_k \triangleq \begin{bmatrix} u_{k,1} & u_{k-1,1} & \cdots & u_{k-N_u+1,1} \\ \vdots & \ldots & \ldots & \vdots \\ u_{k,D_u} & u_{k-1,D_u} & \cdots & u_{k-N_u+1,D_u} \end{bmatrix}.$$

We then reshape both matrices $\bar{y}_{k-1}$ and $\bar{u}_k$ into a single vector $x_k \in \mathbb{R}^{D_x}$, where $D_x = N_y D_y + N_u D_u$:

$$x_k \triangleq \begin{bmatrix} \mathrm{vec}(\bar{y}_{k-1}) \\ \mathrm{vec}(\bar{u}_k) \end{bmatrix}, \tag{1}$$

and $\mathrm{vec}(\cdot)$ denotes the column-wise vectorization operator that stacks the columns of a matrix into a single column vector [26]. At the core of our MARX(N) model is a vector autore-

gressive process with exogenous inputs, characterized by the following likelihood function:

$$\begin{aligned} p(y_k \mid \Theta, x_k) &= \mathcal{N}(y_k \mid A^\intercal x_k, W^{-1}) \\ &= \sqrt{\frac{|W|}{(2\pi)^{D_y}}} \exp\Big(-\frac{1}{2}(y_k - A^\intercal x_k)^\intercal W (y_k - A^\intercal x_k)\Big), \end{aligned} \tag{2}$$

where the parameters—jointly denoted as $\Theta = (A, W)$—consist of a regression coefficient matrix $A \in \mathbb{R}^{D_x \times D_y}$ and a noise precision matrix $W \in \mathbb{R}_+^{D_y \times D_y}$, with $\mathbb{R}_+$ denoting the space of positive semi-definite matrices. Each column $A_{:,j}$ specifies how the full memory vector x_k (comprising past outputs and inputs) linearly predicts the jth component of the current observation $y_{k,j}$. In state–space terminology, A captures both the temporal memory and cross-variable coupling by weighting each lagged signal in x_k. The matrix W represents the inverse covariance (precision) of the Gaussian measurement noise: its diagonal entries set the inverse variances for each observed dimension while off-diagonals model instantaneous noise correlations between different components of y_k.

For computational convenience (see Section 4.1), we specify our prior distribution over Θ as a matrix normal Wishart distribution [27]:

$$p(\Theta) = p(A \mid W)p(W) = \mathcal{MN}(A \mid M_0, \Lambda_0^{-1}, W^{-1})\mathcal{W}(W \mid \Omega_0^{-1}, \nu_0). \tag{3}$$

Here, 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 $W^{-1} \in \mathbb{R}^{D_y \times D_y}$,

$$\begin{aligned} p(A \mid W) &= \mathcal{MN}(A \mid M_0, \Lambda_0^{-1}, W^{-1}) \\ &= \sqrt{\frac{|W|^{D_x}|\Lambda_0|^{D_y}}{(2\pi)^{D_x D_y}}} \exp\Big(-\frac{1}{2}\mathrm{tr}\big[W(A - M_0)^\intercal \Lambda_0 (A - M_0)\big]\Big), \end{aligned} \tag{4}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a square matrix, i.e., the sum of its diagonal entries [26]. The precision matrix W follows a Wishart distribution with a scale matrix $\Omega_0^{-1} \in \mathbb{R}^{D_y \times D_y}$ and degrees of freedom $\nu_0 \in \mathbb{R}$

$$p(W) = \mathcal{W}(W \mid \Omega_0^{-1}, \nu_0) = \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 D_y}}} \frac{\sqrt{|W|^{\nu_0 - D_y - 1}}}{\Gamma_{D_y}(\nu_0/2)} \exp\Big(-\frac{1}{2}\mathrm{tr}\big[W\Omega_0\big]\Big).$$

Here, $\Gamma_{D_y}(\cdot)$ is the multivariate Gamma function with dimension D_y [28]. Our goal is to infer the posterior distribution over A and W and subsequently use these parameter posterior distributions to make predictions for future outputs y_t.

The chosen prior and likelihood define the following generative model over the joint distribution of observations, inputs, and parameters:

$$p(y_{1:k}, u_{1:k}, \Theta) = p(\Theta) \prod_{i=1}^{k} p(y_i \mid \Theta, x_i).$$

We consider two inference paradigms for parameter estimation [29]. In *batch estimation*, the full dataset is used to compute the posterior:

$$p(\Theta \mid y_{1:k}, u_{1:k}) \propto p(\Theta) \prod_{i=1}^{k} p(y_i \mid \Theta, x_i).$$

Alternatively, in *recursive estimation*, the posterior is updated incrementally as new data arrives:

$$p(\Theta \mid y_{1:k}, u_{1:k}) \propto p(\Theta \mid y_{1:k-1}, u_{1:k-1}) p(y_k \mid \Theta, y_{1:k-1}, u_{1:k}) .$$

In this paper, we focus on the recursive formulation, which enables efficient online model updates and is well suited for real-time applications and systems where storing and reprocessing the entire history is infeasible.

Factor Graph

The probabilistic graphical model underlying the recursive formulation is straightforward, consisting of a prior distribution and a likelihood function. Figure 1 presents a Forney-style factor graph in which nodes represent factors, edges denote variables, and each edge connects exactly two nodes [15]. In the graph, time flows from left to right, predictions flow from top to bottom, and corrections flow from bottom to top. The factor node labeled $\mathcal{MNW}$ represents the matrix normal Wishart prediction distribution along with its associated prior parameters. The dashed box represents the composite likelihood node, which comprises (i) the concatenation operation described in (1), (ii) the dot–product operation between the regression coefficient matrix A and the memory x_k, and (iii) the stochastic disturbance. The equality node connects the parameters Θ to the likelihood nodes for each time step k.

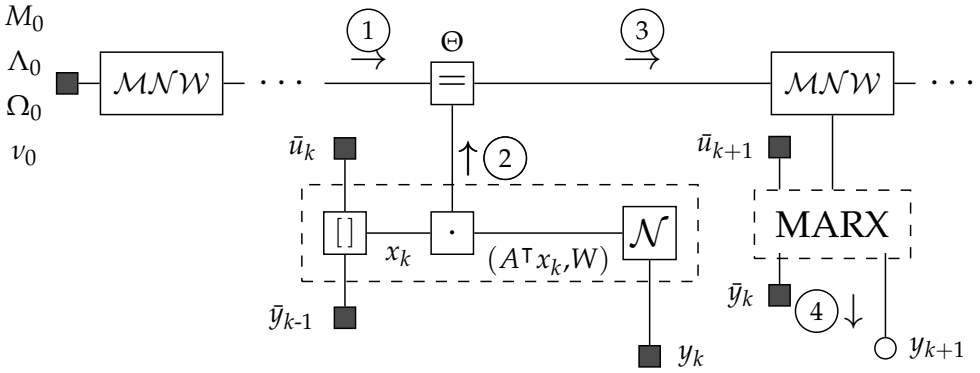


Figure 1. Forney-style factor graph of the MARX model in recursive form. A matrix normal Wishart node sends a prior message (1) to an equality node. A likelihood-based message (2) passes upwards from the MARX likelihood node (dashed box), attached to the observed variables y_k, $\bar{y}_{k-1}$, and $\bar{u}_k$.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

4. Inference

Inference consists of two stages: (i) parameter estimation, where we infer model parameters from observed outputs y_k (Section 4.1), and (ii) output prediction, where we forecast future outputs y_t for $t > k$, given future system inputs u_{k+1} (Section 4.2).

4.1. Parameter Estimation

We wish to recursively estimate the posterior distribution over the model parameters:

$$p(\Theta \mid \mathcal{D}_k) = \frac{p(y_k \mid \Theta, x_k)}{p(y_k \mid u_k, \mathcal{D}_{k-1})} p(\Theta \mid \mathcal{D}_{k-1}) ,$$

where $\mathcal{D}_k = \{y_i, u_i\}_{i=1}^k$ denotes the data up to time k. Note that the memory vector x_k is a subset of $\mathcal{D}_{k-1}$. The evidence term in the denominator is

$$p(y_k \mid u_k, \mathcal{D}_{k-1}) = \int p(y_k \mid \Theta, x_k)\, p(\Theta \mid \mathcal{D}_{k-1})\, \mathrm{d}\Theta\,. \tag{5}$$

This evidence term will be discussed in detail in Section 5.

Lemma 1. *Combining the MARX likelihood (2) with a matrix normal Wishart prior distribution over MARX coefficient matrix A and precision matrix W (3) yields a matrix normal Wishart distribution:*

$$p(\Theta \mid \mathcal{D}_k) = \mathcal{MNW}(A, W \mid M_k, \Lambda_k^{-1}, \Omega_k^{-1}, \nu_k),$$

with the following parameter updates:

$$\begin{aligned}
\nu_k &= \nu_{k-1} + 1 \\
\Lambda_k &= \Lambda_{k-1} + x_k x_k^\intercal \\
M_k &= (\Lambda_{k-1} + x_k x_k^\intercal)^{-1}(\Lambda_{k-1} M_{k-1} + x_k y_k^\intercal) \\
\Omega_k &= \Omega_{k-1} + y_k y_k^\intercal + M_{k-1}^\intercal \Lambda_{k-1} M_{k-1} \\
&\quad - (\Lambda_{k-1} M_{k-1} + x_k y_k^\intercal)^\intercal (\Lambda_{k-1} + x_k x_k^\intercal)^{-1} (\Lambda_{k-1} M_{k-1} + x_k y_k^\intercal)\,.
\end{aligned}$$

See Appendix A for the proof. This solution can be cast as a message-passing procedure on a factor graph, allowing distributed computation [15,30].

In Figure 1, circled messages indicate the information flow between the factor nodes along the edges. Message ① represents the previous posterior belief over $\Theta = (A, W)$:

$$\overrightarrow{①} = p(\Theta \mid \mathcal{D}_{k-1}) = \mathcal{MNW}(A, W \mid M_{k-1}, \Lambda_{k-1}^{-1}, \Omega_{k-1}^{-1}, \nu_{k-1})\,. \tag{6}$$

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

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

$$\uparrow ② = p(y_k \mid \Theta, x_k) \propto \mathcal{MNW}(A, W \mid \bar{M}_k, \bar{\Lambda}_k^{-1}, \bar{\Omega}_k^{-1}, \bar{\nu}_k)\,. \tag{7}$$

Its parameters are

$$\begin{aligned}
\bar{\nu}_k &= 2 - D_x + D_y\,, & \bar{\Lambda}_k &= x_k x_k^\intercal\,, \\
\bar{M}_k &= (x_k x_k^\intercal)^{-1} x_k y_k^\intercal\,, & \bar{\Omega}_k &= 0_{D_y \times D_y}\,.
\end{aligned}$$

See Appendix B for the proof. Note that the scale matrix is not positive-definite, which implies that message ② is an improper distribution. Utilizing improper distributions is not uncommon when messages are intermediate results. For example, in variational and particle-based message passing, the messages are unnormalized and therefore also technically improper distributions [31,32]. However, should one want to visualize message ② or convert it to a related distribution, for instance, then the scale matrix can be perturbed with a machine precision offset (i.e., $\bar{\Omega}_k = 10^{-8} \cdot I_{D_y \times D_y}$).

Message ③ results from multiplying messages ① and ② at the equality node [15].

Lemma 3. *Let p_1 and p_2 be two matrix normal Wishart distributions over the same random variables Θ:*

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

Their product is proportional to another matrix normal Wishart distribution:

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

and its parameters are combinations of p_1, p_2's parameters,

$$\begin{aligned}
\nu_3 &= \nu_1 + \nu_2 + D_x - D_y - 1\,,\\
\Lambda_3 &= \Lambda_1 + \Lambda_2\,,\\
M_3 &= (\Lambda_1 + \Lambda_2)^{-1}(\Lambda_1 M_1 + \Lambda_2 M_2)\,,\\
\Omega_3 &= \Omega_1 + \Omega_2 + M_1^\intercal \Lambda_1 M_1 + M_2^\intercal \Lambda_2 M_2\\
&\quad - (\Lambda_1 M_1 + \Lambda_2 M_2)^\intercal (\Lambda_1 + \Lambda_2)^{-1}(\Lambda_1 M_1 + \Lambda_2 M_2).
\end{aligned}$$

See Appendix C for the proof.

Theorem 1. *The outgoing message from the equality node is proportional to the exact recursive posterior distribution:*

$$\overrightarrow{③} = \overrightarrow{①} \cdot ② \uparrow \propto \mathcal{MNW}(A, W \mid M_k, \Lambda_k^{-1}, \Omega_k^{-1}, \nu_k).$$

Proof. Combining parameters from the messages in (6) and (7) according to the product operation in Lemma 3 yields

$$\begin{aligned}
\nu_k &= \nu_{k-1} + \bar{\nu}_k + D_x - D_y - 1 = \nu_{k-1} + 1\,,\\
\Lambda_k &= \Lambda_{k-1} + \bar{\Lambda}_k = \Lambda_{k-1} + x_k x_k^\intercal\,,\\
M_k &= (\Lambda_{k-1} + \bar{\Lambda}_k)^{-1}(\Lambda_{k-1} M_{k-1} + \bar{\Lambda}_k \bar{M}_k) = (\Lambda_{k-1} + x_k x_k^\intercal)^{-1}(\Lambda_{k-1} M_{k-1} + x_k y_k^\intercal)\,,\\
\Omega_k &= \Omega_{k-1} + \bar{\Omega}_k + M_{k-1}^\intercal \Lambda_{k-1} M_{k-1} + \bar{M}_k^\intercal \bar{\Lambda}_k \bar{M}_k\\
&\quad - (\Lambda_{k-1} M_{k-1} + \bar{\Lambda}_k \bar{M}_k)^\intercal (\Lambda_{k-1} + \bar{\Lambda}_k)^{-1}(\Lambda_{k-1} M_{k-1} + \bar{\Lambda}_k \bar{M}_k)\\
&= \Omega_{k-1} + M_{k-1}^\intercal \Lambda_{k-1} M_{k-1} + y_k y_k^\intercal\\
&\quad - (\Lambda_{k-1} M_{k-1} + x_k y_k^\intercal)^\intercal (\Lambda_{k-1} + x_k x_k^\intercal)^{-1}(\Lambda_{k-1} M_{k-1} + x_k y_k^\intercal).
\end{aligned}$$

These match the parameter update rules outlined in Lemma 1. □

4.2. Output Prediction

Predicting future system outputs amounts to computing the posterior predictive distribution, i.e., the marginal distribution of y_t for $t > k$:

$$\downarrow ④ = p(y_t \mid u_t, \mathcal{D}_k) = \int p(y_t \mid \Theta, x_t) p(\Theta \mid \mathcal{D}_k)\, \mathrm{d}\Theta\,. \tag{8}$$

We exploit the factorization of the parameter posterior over (A, W) to split this into a marginalization over A:

$$p(y_t \mid W, u_t, \mathcal{D}_k) = \int p(y_t \mid \Theta, x_t) p(A \mid W, \mathcal{D}_k)\, \mathrm{d}A\,,$$

and a marginalization over W:

$$p(y_t \mid u_t, \mathcal{D}_k) = \int p(y_t \mid W, u_t, \mathcal{D}_k) p(W \mid \mathcal{D}_k) \, \mathrm{d}W \, .$$

Theorem 2. *Marginalizing the composite MARX likelihood (2) over the matrix normal distribution (4) for A yields a multivariate normal distribution:*

$$\int \mathcal{N}\left(y_t \mid A^\intercal x_t, W^{-1}\right) \mathcal{MN}\left(A \mid M_k, \Lambda_k^{-1}, W^{-1}\right) \mathrm{d}A = \mathcal{N}\left(y_t \mid M_k^\intercal x_t, (\lambda_t W)^{-1}\right) ,$$

where $\lambda_t \triangleq (1 + x_t^\intercal \Lambda_k^{-1} x_t)^{-1}$.

See Appendix D for the proof.

Theorem 3. *Marginalizing a multivariate normal distribution over a Wishart distribution on its precision parameter yields a multivariate location-scale Student's t-distribution [27]:*

$$\int \mathcal{N}\left(y_t \mid M_k^\intercal x_t, (\lambda_t W)^{-1}\right) \mathcal{W}(W \mid \Omega_k^{-1}, \nu_k) \, \mathrm{d}W = \mathcal{T}(y_t \mid \mu_t, \Psi_t^{-1}, \eta_t) \, , \tag{9}$$

where $\mu_t \triangleq M_k^\intercal x_t$, $\eta_t \triangleq \nu_k - D_y + 1$, *and* $\Psi_t \triangleq \eta_t \Omega_k^{-1} \lambda_t$.

See Appendix E for the proof. The resulting posterior predictive distribution provides a recursive estimate of output uncertainty, which is valuable for decision-making and adaptive control.

5. Model Evaluation

A key criterion for probabilistic model evaluation is the negative log-model evidence (or surprisal) $-\log p(y_k)$, which quantifies how surprising the observed data y_k is under the model [33,34]. To gain deeper insights into model performance, we analyze surprisal from the perspective of variational inference on factor graphs. This approach enables us to decompose the overall model score into contributions from the individual nodes and edges of the graph.

Variational inference casts Bayesian inference as an optimization problem by approximating the true posterior $p(\Theta \mid \mathcal{D}_k)$ with a computationally tractable variational posterior $q(\Theta \mid \mathcal{D}_k)$, chosen from a variational family Q [33,35]. At time k, the optimal variational posterior is obtained by minimizing variational free energy (VFE) [36,37]:

$$q^*(\Theta \mid \mathcal{D}_k) = \arg\min_{q \in Q} \mathcal{F}_{VFE}\Big[q(\Theta \mid \mathcal{D}_k), p(y_k, \Theta) \Big] \, ,$$

where the VFE functional $\mathcal{F}_{VFE}$ is defined as

$$\mathcal{F}_{VFE}\Big[q(\Theta \mid \mathcal{D}_k), p(y_k, \Theta)\Big] = \underbrace{D_{KL}[q(\Theta \mid \mathcal{D}_k) \mid\mid p(\Theta \mid \mathcal{D}_k)]}_{\text{Inference Cost}} - \log \underbrace{p(y_k \mid u_k, \mathcal{D}_{k-1})}_{\text{Model Evidence}} \, .$$

In exact inference, where the true posterior is computed via Bayes' rule, the inference cost becomes zero, and the VFE equals the exact surprisal. When exact inference is intractable, VFE is expressed in a different way. By absorbing the evidence term into the Kullback–Leibler (KL)-divergence, the product of the posterior and the evidence becomes the joint distribution of the generative model, which can be decomposed into a likelihood

times prior distribution. This yields the decomposition of free energy into complexity and accuracy terms [37]:

$$\begin{aligned} & D_{KL}[q(\Theta \mid \mathcal{D}_k) || p(\Theta \mid \mathcal{D}_k)] - \log p(y_k \mid u_k, \mathcal{D}_{k-1}) \\ & \quad = \mathbb{E}_{q(\Theta \mid \mathcal{D}_k)}\big[\log \frac{q(\Theta \mid \mathcal{D}_k)}{p(y_k, \Theta \mid \mathcal{D}_k)}\big] \\ & \quad = \underbrace{D_{KL}[q(\Theta \mid \mathcal{D}_k) \;||\; p(\Theta \mid \mathcal{D}_{k-1})]}_{\text{Complexity}} + \underbrace{H[q(\Theta \mid \mathcal{D}_k), p(y_k \mid \Theta, x_k)]}_{\text{Accuracy}}, \end{aligned} \tag{10}$$

where complexity measures how much the variational posterior deviates from the prior, penalizing unnecessary deviations from prior knowledge and controlling overfitting. Accuracy quantifies the model's ability to explain the observed data, expressed as the expected negative log-likelihood under the variational posterior. To refine this decomposition further, we introduce an auxiliary entropy term $H(\Theta \mid \mathcal{D}_k)$ and rewrite (10) as

$$\begin{aligned} & \mathcal{F}_{VFE}\Big[q(\Theta \mid \mathcal{D}_k), p(y_k, \Theta)\Big] \\ & = D_{KL}[q(\Theta|\mathcal{D}_k) \| p(\Theta|\mathcal{D}_{k\text{-}1})] + H[q(\Theta|\mathcal{D}_k), p(y_k|\Theta, x_k))] - H[q(\Theta|\mathcal{D}_k)] + H[q(\Theta|\mathcal{D}_k)] \\ & = D_{KL}[q(\Theta \mid \mathcal{D}_k) \;||\; p(\Theta \mid \mathcal{D}_{k-1})] + D_{KL}[q(\Theta \mid \mathcal{D}_k) \;||\; p(y_k \mid \Theta, x_k))] + H[q(\Theta \mid \mathcal{D}_k)]. \end{aligned} \tag{11}$$

For models formulated as Forney-style factor graphs, inference is performed by optimizing the Bethe Free Energy (BFE), a generalization of VFE, which accounts for the graph's structure [13,21,38]:

$$\mathcal{F}_{BFE}[\, q(\Theta \mid \mathcal{D}_k), p(y_k, \Theta)\,] \triangleq \sum_{a \in \mathcal{V}} D_{KL}[q_a || p_a] + \sum_{i \in \mathcal{E}} H[q_i], \tag{12}$$

where $\mathcal{V}$ is the set of factor nodes and $\mathcal{E}$ is the set of edges. In this formulation, each q_a is the local variational belief at node a, p_a is the corresponding exact local distribution, and each edge i contributes an entropy term $H[q_i]$. In our recursive MARX model—comprising a MARX likelihood node, a prior node, and an edge for the joint parameters Θ—the BFE decomposition in (12) coincides with the VFE decomposition in (11). Thus, factor graphs enable a fine-grained attribution of surprisal to specific components of the system.

5.1. MARX Model Evidence and Surprisal

To evaluate the model properly, we must compute the model evidence (marginal likelihood), which is the probability of an observed sample marginalized over parameters, weighted by their prior probabilities. Equation (5) already detailed the evidence term, but this still involved an integral. This integral is identical to the integral for the posterior predictive distribution (8), except that y_k and u_k are observed and the prior parameters are those from time step $k-1$. Concretely,

$$\begin{aligned} & p(y_k \mid u_k, \mathcal{D}_{k-1}) = \int p(y_k \mid \Theta, x_k) p(\Theta \mid \mathcal{D}_{k-1})\, \mathrm{d}\Theta = \mathcal{T}(y_k \mid m_k, \Psi_k^{-1}, \eta_k) \\ & = \sqrt{\frac{|\Psi_k|}{(\eta_k \pi)^{D_y}}} \frac{\Gamma_{D_y}((\eta_k + D_y)/2)}{\Gamma_{D_y}((\eta_k + D_y - 1)/2)} \big(1 + \frac{1}{\eta_k}(y_k - m_k)^\intercal \Psi_k (y_k - m_k)\big)^{-(\eta_k + D_y)/2}, \end{aligned}$$

where $m_k = M_{k-1}^\intercal x_k$, $\eta_k = \nu_{k-1} - D_y + 1$, $\Psi_k = \eta_k \Omega_{k-1}^{-1} \lambda_k$, and $\lambda_k = (1 + x_k^\intercal \Lambda_{k-1}^{-1} x_k)^{-1}$. Here $\mathcal{T}(\cdot \mid \mu, \Sigma^{-1}, \nu)$ denotes the multivariate Student's t-distribution with location μ, scale Σ^{-1}, and degrees of freedom ν. Unlike the posterior predictive distribution, the model

evidence is a scalar: higher values indicate that the model better explains the observed data. Hence, the surprisal for our model is

$$\begin{aligned}-\log p(y_k \mid u_k, \mathcal{D}_{k-1}) = &-\frac{1}{2}\log|\Psi_k| + \frac{D_y}{2}\log(\eta_k\pi) \\ &- \log\Gamma_{D_y}(\frac{\eta_k + D_y}{2}) + \log\Gamma_{D_y}(\frac{\eta_k + D_y - 1}{2}) \\ &+ \frac{\eta_k + D_y}{2}\log\left(1 + \frac{1}{\eta_k}(y_k - m_k)^\intercal \Psi_k (y_k - m_k)\right). \end{aligned} \tag{13}$$

5.2. MARX Variational Free Energy

Lemma 4. *Let q and p be two matrix normal Wishart distributions over the same random variables Θ, representing the posterior and prior, respectively:*

$$\begin{aligned} q(\Theta \mid \mathcal{D}_k) &= \mathcal{MNW}(\Theta \mid M_k, \Lambda_k^{-1}, \Omega_k^{-1}, \nu_k) \\ p(\Theta \mid \mathcal{D}_{k-1}) &= \mathcal{MNW}(\Theta \mid M_{k-1}, \Lambda_{k-1}^{-1}, \Omega_{k-1}^{-1}, \nu_{k-1}). \end{aligned}$$

The differential cross-entropy $H[q(\Theta \mid \mathcal{D}_k), p(\Theta \mid \mathcal{D}_{k-1})]$ of the posterior relative to the prior is

$$\begin{aligned} H[q(\Theta \mid \mathcal{D}_k), p(\Theta \mid \mathcal{D}_{k-1})] = &-\frac{1}{2}D_y \log|\Lambda_{k-1}| + \frac{1}{2}(\nu_{k-1} + D_x - D_y - 1)\log|\Omega_k| \\ &- \frac{1}{2}\nu_{k-1}\log|\Omega_{k-1}| + \frac{1}{2}(D_y + 1)D_y \log 2 + \frac{1}{2}D_x D_y \log\pi \\ &+ \log\Gamma_{D_y}(\frac{\nu_{k-1}}{2}) - \frac{1}{2}(\nu_{k-1} + D_x - D_y - 1)\psi_{D_y}(\frac{\nu_k}{2}) \\ &+ \frac{1}{2}\nu_k tr\left(\Omega_k^{-1}(M_k - M_{k-1})^\intercal \Lambda_{k-1}(M_k - M_{k-1})\right) \\ &+ \frac{1}{2}D_y tr(\Lambda_k^{-1}\Lambda_{k-1}^\intercal) + \nu_k tr(\Omega_k^{-1}\Omega_{k-1}). \end{aligned}$$

See Appendix F for the proof.

Lemma 5. *Consider the matrix normal Wishart posterior:*

$$q(\Theta \mid \mathcal{D}_k) = \mathcal{MNW}(A, W \mid M_k, \Lambda_k^{-1}, \Omega_k^{-1}, \nu_k).$$

Its (differential) entropy is

$$\begin{aligned} H[q(\Theta \mid \mathcal{D}_k)] = &-\frac{1}{2}D_y \log|\Lambda_k| + \frac{1}{2}(D_x - D_y - 1)\log|\Omega_k| \\ &+ \frac{1}{2}(D_y + 1)D_y \log 2 + \frac{1}{2}D_x D_y \log\pi + \frac{1}{2}(D_x + \nu_k)D_y \\ &+ \log\Gamma_{D_y}(\frac{\nu_k}{2}) - \frac{1}{2}(\nu_k + D_x - D_y - 1)\psi_{D_y}(\frac{\nu_k}{2}). \end{aligned} \tag{14}$$

See Appendix G for the proof.

Lemma 6. *Let q and p be two matrix normal Wishart distributions over the same random variables Θ, representing the posterior and prior, respectively:*

$$\begin{aligned} q(\Theta \mid \mathcal{D}_k) &= \mathcal{MNW}(\Theta \mid M_k, \Lambda_k^{-1}, \Omega_k^{-1}, \nu_k) \\ p(\Theta \mid \mathcal{D}_{k-1}) &= \mathcal{MNW}(\Theta \mid M_{k-1}, \Lambda_{k-1}^{-1}, \Omega_{k-1}^{-1}, \nu_{k-1}). \end{aligned}$$

The KL-divergence $D_{KL}[q(\Theta \mid \mathcal{D}_k) \mid\mid p(\Theta \mid \mathcal{D}_{k-1})]$ *of the posterior from the prior (complexity) is*

$$\begin{aligned} D_{KL}[q(\Theta \mid \mathcal{D}_k) \mid\mid p(\Theta \mid \mathcal{D}_{k-1})] = \frac{1}{2} D_y \log \frac{|\Lambda_k|}{|\Lambda_{k-1}|} + \frac{1}{2} \nu_{k-1} \log \frac{|\Omega_k|}{|\Omega_{k-1}|} - \frac{1}{2}(D_x + \nu_k) D_y \\ - \log \Gamma_{D_y}(\frac{\nu_k}{2}) + \log \Gamma_{D_y}(\frac{\nu_{k-1}}{2}) + \frac{1}{2}(\nu_k - \nu_{k-1}) \psi_{D_y}(\frac{\nu_k}{2}) \\ + \frac{1}{2} \nu_k tr\Big(\Omega_k^{-1}(M_k - M_{k-1})^\intercal \Lambda_{k-1}(M_k - M_{k-1})\Big) \\ + \frac{1}{2} D_y tr(\Lambda_k^{-1} \Lambda_{k-1}^\intercal) + \nu_k tr(\Omega_k^{-1} \Omega_{k-1}) \,. \end{aligned}$$

See Appendix H for the proof.

Lemma 7. *Consider a matrix normal Wishart distribution q and a multivariate normal distribution p, representing the posterior and MARX likelihood:*

$$\begin{aligned} q(\Theta \mid \mathcal{D}_k) &= \mathcal{MNW}(A, W \mid M_k, \Lambda_k^{-1}, \Omega_k^{-1}, \nu_k) \\ p(y_k \mid \Theta, x_k) &= \mathcal{N}(y_k \mid A^\intercal x_k, W^{-1}) \,. \end{aligned}$$

The differential cross-entropy $H[q(\Theta \mid \mathcal{D}_k), p(y_k \mid \Theta, x_k)]$ *of the posterior relative to the likelihood (accuracy) is*

$$\begin{aligned} H[q(\Theta \mid \mathcal{D}_k), p(y_k \mid \Theta, x_k)] = -\frac{1}{2} \psi_{D_y}(\frac{\nu_k}{2}) + \frac{1}{2} \log |\Omega_k| + \frac{1}{2} D_y \log \pi \\ + \frac{1}{2} \nu_k (y_k - M_k^\intercal x_k)^\intercal \Omega_k^{-1} (y_k - M_k^\intercal x_k) + \frac{1}{2} x_k^\intercal \Lambda_k^{-1} x_k D_y \,. \end{aligned}$$

See Appendix I for the proof.

6. Experiments

We conducted three experiments: one verification experiment and two validation experiments (Code: https://github.com/biaslab/MDPI2025-MARX, accessed on 8 March 2025). In the verification experiment (Section 6.2), we tested whether the MARX estimator could identify a dynamical system with known parameters. In the validation experiments (Section 6.3), we assess the estimator's performance on two complex dynamical systems with unknown parameters: a linear double mass-spring-damper system and a nonlinear double pendulum. In all the experiments, we compare the performance of the MARX estimator to a baseline approach.

6.1. Baseline Estimator

We compare against a recursive least squares (RLS) estimator [3]. Let $\hat{A}_k$ be a point estimate of the coefficient matrix based on the previous k data points, and let $P_0 = I_{D_x}$ be an initial inverse sample covariance matrix. These matrices are updated at each time step according to

$$\begin{aligned} P_k &= P_{k-1} - P_{k-1} x_k (1 + x_k^\intercal P_{k-1} x_k)^{-1} x_k^\intercal P_{k-1} \\ \hat{A}_k &= \hat{A}_{k-1} + P_{k-1} x_k (1 + x_k^\intercal P_{k-1} x_k)^{-1} (y_k - \hat{A}_{k-1}^\intercal x_k)^\intercal \,. \end{aligned}$$

Note that this formulation corresponds to a forgetting factor of 1.0, meaning that older data points are not down-weighted. The system outputs are predicted with $y_t = \hat{A}_k^\intercal x_t$.

6.2. Verification

We perform a verification experiment on a MARX system with state $z_k = x_k$ (1), memory sizes $N_y = 2, N_u = 3$, and dimensions $D_y = D_u = 2$. The system has true parameters $\tilde{\Theta} = (\tilde{A}, \tilde{W})$. It evolves according to $g(f(x_k)) = \tilde{A}^\intercal x_k$, where $\tilde{A}$ is the known coefficient matrix (see Figure 2). For each output dimension i, the lag-dependent coefficients were generated using a Butterworth low-pass filter (cutoff frequency 20 Hz) applied to that same dimension, while cross-dimensional coefficients were sampled from $\mathcal{N}(0, 0.1^2)$ [39]. We chose the Butterworth filter because its maximally flat response in the passband ensures that signals below the cutoff frequency are transmitted with little distortion while attenuating higher-frequency components [40]. This makes it suitable for generating stable linear dynamics and mimicking the low-pass behavior often observed in physical dynamical systems—such as mechanical or electrical processes [41,42]—and is common in applications like audio and biomedical signal processing [41,43]. The disturbance follows $e_k \sim \mathcal{N}(0, \tilde{W}^{-1})$ with precision matrix $\tilde{W} = \begin{bmatrix} 300 & 100 \\ 100 & 200 \end{bmatrix}$.

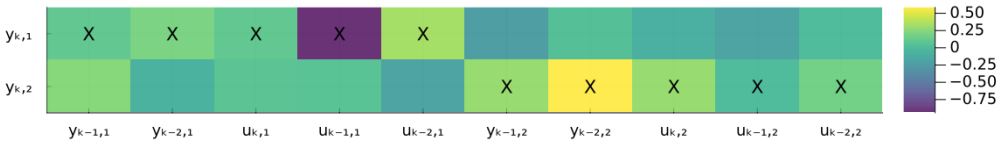


Figure 2. Heatmap of true system parameter $\tilde{A}^\intercal$. "X" denotes coefficients generated from a Butterworth filter.

We evaluated each estimator for training sizes $T_{\text{train}} \in \{2^l \mid l \in \{2,3,4,5,6\}\}$, using Monte Carlo experiments with $N_{MC} = 100$ runs. To learn the parameters, each estimator uses T_{train} state transitions, starting from state $z_0 = 0_{D_z}$. After training, each estimator is tested for $T_{\text{test}} = 100$ time steps, again starting from z_0 but with different control signals. For the MARX estimator, we compare two priors (see Table 1): uninformative (MARX-UI) and weakly informative (MARX-WI). The uninformative prior uses small precision values for Λ_0 and Ω_0, corresponding to large prior variancesthat reflect minimal prior belief about the parameters. The weakly informative prior assigns higher precision (lower variance), introducing a mild preference for more stable parameter values while still letting the data dominate. In both cases, the degrees of freedom ν_0 are kept minimal at $D_y + 3$, just above the threshold for the Wishart distribution to be well defined, further reinforcing the limited informativeness of the prior. The weakly informative prior also encodes approximate prior knowledge about the observation noise. Specifically, the Wishart component $p(W)$ has a mode at $\nu_0 \Omega_0^{-1} = \begin{bmatrix} 500 & 0 \\ 0 & 500 \end{bmatrix}$, which is of similar magnitude to the true noise precision $\tilde{W}$. In contrast, the uninformative prior sets Ω_0 to much larger, placing its mode far from the true noise characteristics. Thus, the weakly informative prior softly incorporates domain knowledge about expected noise levels, improving convergence and stability in the early stages of recursive estimation. For each training size, we calculate the root mean squared error (RMSE),

$$\text{RMSE} = \sqrt{\frac{1}{T_{\text{test}}} \sum_{k=1}^{T_{\text{test}}} (\hat{y}_k - y_k)^2},$$

between the predicted output $\hat{y}_k$, i.e., the mean of the posterior predictive $p(y_k \mid u_k, \mathcal{D}_{k-1})$, and the true output y_k for all $k \in T_{\text{test}}$ evaluation steps.

Table 1. Sets of prior parameters used in the experiments.

	M_0	Λ_0	Ω_0	ν_0
Uninformative	$0_{D_x \times D_y}$	$1 \times 10^{-4} \cdot I_{D_x}$	$1 \times 10^{-5} \cdot I_{D_y}$	$D_y + 3$
Weakly informative	$0_{D_x \times D_y}$	$1 \times 10^{-1} \cdot I_{D_x}$	$1 \times 10^{-2} \cdot I_{D_y}$	$D_y + 3$

Figure 3 shows the simulation errors for MARX-UI, MARX-WI, and RLS as a function of the training size. For small sample sizes, MARX-WI consistently outperforms RLS, while MARX-UI performs slightly worse. All three estimators converge to the same performance level as the training size increases.

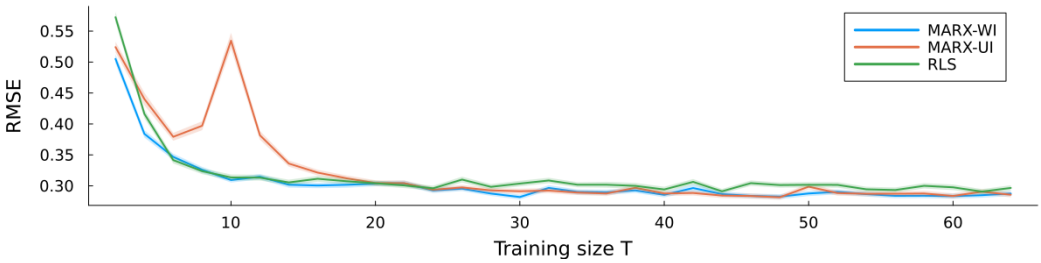


Figure 3. Simulation errors (average RMSE) of all three estimators for the MARX system, with ribbons indicating standard errors.

Figure 4 focuses on a single Monte Carlo experiment with $T_{\text{train}} = 2^6$. It plots $\log(||\tilde{A} - A||_F)$, the log of the Frobenius norm between the true coefficient matrix $\tilde{A}$ and each estimate A. MARX-WI consistently yields better estimates of $\tilde{A}$ than MARX-UI and RLS. Although MARX-UI struggles during the first 25 time steps, it eventually produces a more accurate estimate of $\tilde{A}$ compared to RLS.

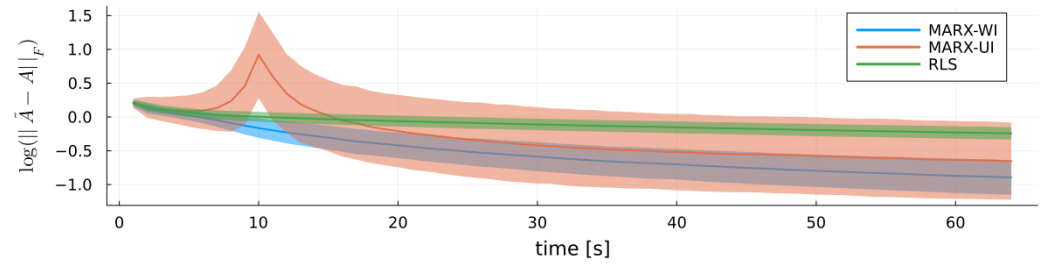


Figure 4. Log-scale Frobenius norm of the difference between true coefficient matrix $\tilde{A}$ and estimates A of each estimator in a single Monte Carlo run with $T_{\text{train}} = 2^6$ for the MARX system, with ribbons indicating standard errors.

Unlike RLS, the MARX estimator also estimates the noise precision matrix W. Figure 5 shows $\log(||\tilde{W} - W||_F)$ for both MARX-WI and MARX-UI. MARX-WI consistently achieves more accurate estimates of $\tilde{W}$ than MARX-UI.

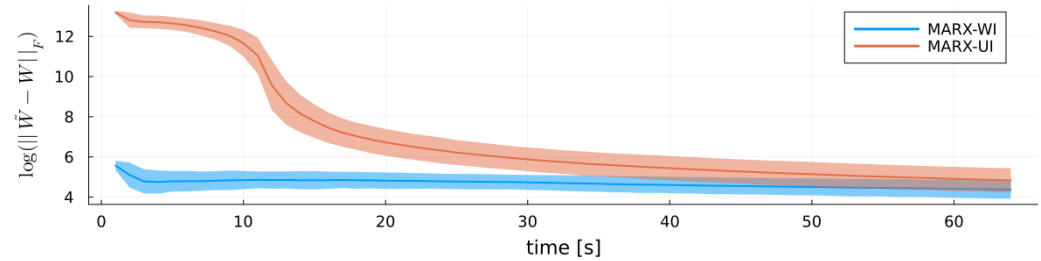


Figure 5. Log-scale Frobenius norm of the difference between true coefficient matrix $\tilde{W}$ and estimates W of each MARX estimator for the MARX system, with ribbons indicating standard errors.

Figure 6 plots the negative log posterior probability of the true parameters $\tilde{\Theta}$ (lower is better), showing that the posterior concentrates sharply on the true values. As a probabilistic estimator, MARX also quantifies uncertainty in its estimates of $\tilde{A}$ and $\tilde{W}$ via the posterior precision (or scale) parameters. Figure 7 illustrates the evolution of MARX-WI's estimates of W for a single run with $T_{\text{train}} = 2^6$. The ribbon represents one standard deviation around the mean. Initially, MARX-WI exhibits high uncertainty (large variance), which generally decreases over time. Because $\tilde{W}$ and W are symmetric, only the upper-triangular elements are shown.

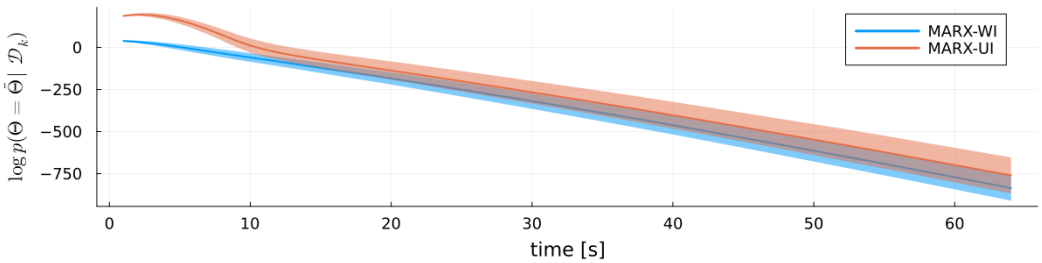


Figure 6. Negative log posterior probability of the true system parameters $\tilde{\Theta}$ under each prior choice for the MARX system (lower is better).

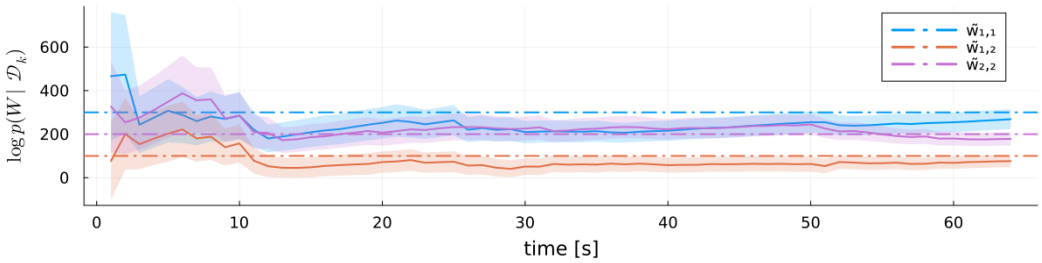


Figure 7. Time series of the estimated noise precision matrix W for the MARX-WI for the MARX system. Ribbons indicate one standard deviation, and horizontal lines denote the true values of $\tilde{W}$.

Figure 8 (top) shows a heatmap of the difference $A - \tilde{A}$. To save space, we plot only a subset of the elements of A, marked by "X". This subset includes the elements with the largest estimation errors and two randomly selected elements. Figure 8 (bottom) shows the evolution of these selected elements for the same Monte Carlo experiment run, with ribbons indicating one standard deviation around each mean estimate.

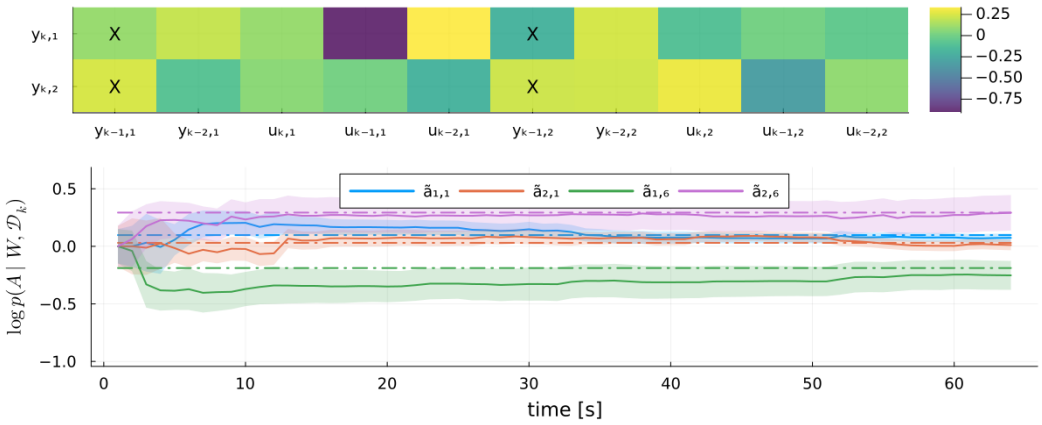


Figure 8. Top: Heatmap of the final $\tilde{A}^\intercal$ coefficient matrix parameter estimate by the MARX-WI model. "X" marks selected elements, and the trajectories are shown below. **Bottom**: Time series of the selected elements of $\tilde{A}$ estimated by MARX-WI , with ribbons indicating one standard deviation. Horizontal lines show the true values of the corresponding elements of $\tilde{A}$.

Furthermore, we apply the model score decomposition from Section 5 to evaluate our recursive MARX model. By tracking how surprisal and its constituent terms evolve, we obtain fine-grained insights into the model's learning dynamics and uncertainty reduction. We can recall from (10) that surprisal decomposes into an accuracy term—given by the cross-entropy of the variational posterior relative to the likelihood, reflecting data fit—and a complexity term—given by the KL-divergence of the variational posterior from the prior, quantifying deviation from prior beliefs. Figure 9 illustrates this decomposition. In the early stages of model training, the complexity term (green) dominates overall surprisal (dashed blue), indicating substantial updates from the prior as the model learns the system parameters. As training progresses and the posterior stabilizes, the complexity term diminishes, and the accuracy term (red) becomes the main source of uncertainty. Spikes in overall surprisal during later stages align with spikes in the accuracy term, which we interpret as indicators of measurement outliers that temporarily degrade model fit.

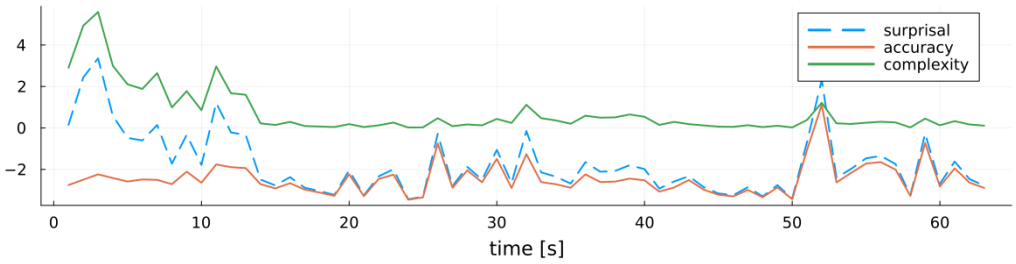


Figure 9. MARX-WI surprisal (dashed blue line) and its decomposition into accuracy (red line) and complexity (green line) over time for the MARX system.

Figure 10 complements this analysis by plotting the entropy of the variational posterior $q(\Theta \mid \mathcal{D}_k)$ over time. This highlights how quickly the inference procedure narrows the parameter space, providing insight into convergence speed and residual uncertainty in the model parameters.

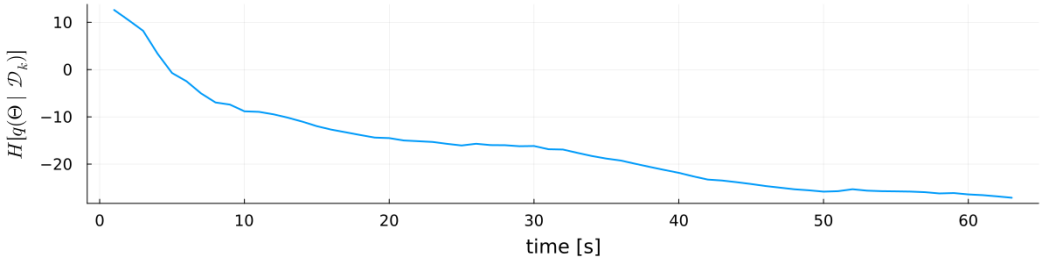


Figure 10. Entropy of the MARX-WI variational posterior $q(\Theta \mid \mathcal{D}_k)$ over time for the MARX system.

We also demonstrate model evaluation using model evidence. Figure 11 shows the evolution of surprisal (lower is better) over time for MARX-WI and MARX-UI. This plot highlights that the prior choice matters only initially; with sufficient data, MARX-WI and MARX-UI converge to the same performance.

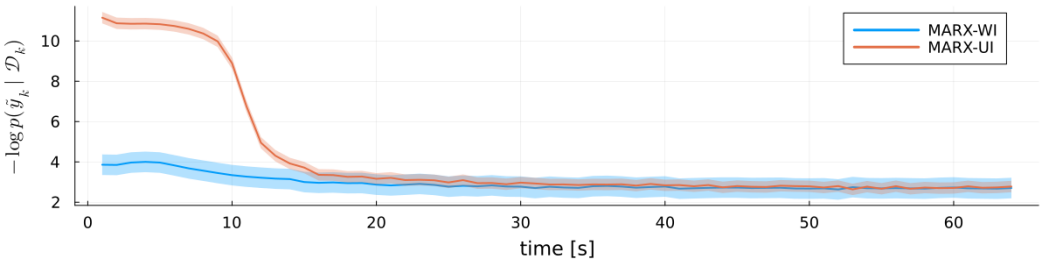


Figure 11. Surprisal over time for MARX-WI versus MARX-UI for the MARX system.

6.3. Validation

To evaluate the proposed method, we perform validation experiments on two distinct mechanical systems: a *linear* double mass-spring-damper system and a *nonlinear* double pendulum system. These testbeds span a range of dynamical complexity and are standard benchmarks for modeling and control tasks. Despite their differences, both systems share a common formulation as second-order dynamical systems expressed in first-order ODE form:

$$I_k \ddot{z}_k = F(z_k, \dot{z}_k, u_k)\,,$$

where z_k denotes generalized coordinates, $\dot{z}_k$ and $\ddot{z}_k$ are the first and second time derivatives of z_k, u_k are the control inputs, I_k is a (state-dependent) generalized inertia matrix, and F encodes the system-specific generalized forces (including passive dynamics and external control inputs). Time evolution is performed using a forward Euler integrator with a system-specific time step Δt:

$$z_{k+1} = z_k + \Delta t \dot{z}_k \quad \text{and} \quad \dot{z}_{k+1} = \dot{z}_k + \Delta t \ddot{z}_k\,.$$

For both validation systems, we choose a disturbance $e_k \sim \mathcal{N}(0, \tilde{W}^{-1})$ with a precision matrix $\tilde{W} = \begin{bmatrix} 2000 & 1000 \\ 1000 & 2000 \end{bmatrix}$. The validation experiments follow the same procedure as the verification experiment: we perform Monte Carlo experiments with $N_{MC} = 100$ runs with $\Delta t = 0.05$, in which each estimator has $T_{\text{train}} \in \{2^l \mid l \in \{2, 3, 4, 5, 6\}\}$ state transitions to learn the parameters (starting from state $z_0 = 0_{D_z}$), and we test each estimator with $T_{\text{test}} = 100$ transitions. However, we increase the memory sizes of the MARX model to $N_y = N_u = 5$.

In the following, we describe each validation system individually, and then present the combined validation results.

6.3.1. Linear System: Double Mass-Spring-Damper

The linear system consists of two masses: $m_1 = 1.0\,\text{kg}$, connected to a fixed base by a spring and damper with stiffness $k_1 = 0.99$ and damping $c_1 = 0.4$, and $m_2 = 2.0\,\text{kg}$, connected to m_1 via a second spring and damper with $k_2 = 0.8$ and $c_2 = 0.4$. The generalized coordinates $z_k \in \mathbb{R}^2$ represent the displacements of each mass from the equilibrium, and the generalized inertia matrix is a constant: $I_k = \text{diag}(m_1, m_2)$, where $\text{diag}(\cdot)$ denotes a diagonal matrix with the given entries [26]. The generalized force function F combines the internal spring and damping forces with external inputs:

$$F(z_k, \dot{z}_k, u_k) = Kz_k + C\dot{z}_k + u_k\,,$$

with the stiffness and damping matrices:

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

6.3.2. Nonlinear System: Double Pendulum

The nonlinear system is a planar double pendulum (also called an acrobot) with two links of lengths $l_1 = 1.0\,\text{m}$ and $l_2 = 1.0\,\text{m}$ and masses $m_1 = 1.0\,\text{kg}$ and $m_2 = 1.0\,\text{kg}$, respectively. The generalized coordinates $z_k \in \mathbb{R}^2$ represent the joint angles, and the generalized inertia matrix is captured implicitly through a structured nonlinear force

formulation. The dynamics are governed by gravity and nonlinear velocity coupling, yielding

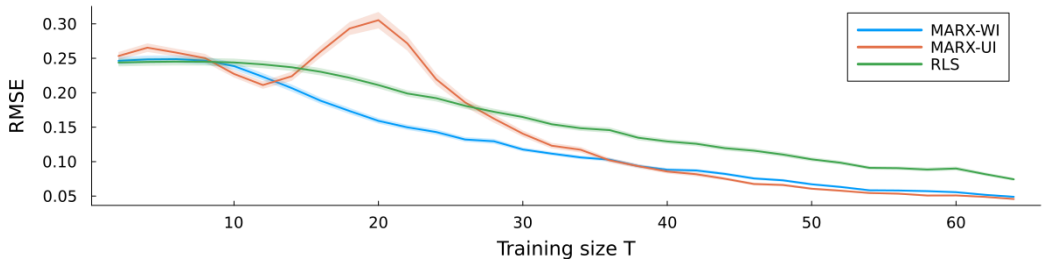
$$F(z_k, \dot{z}_k, u_k) = \operatorname{diag}\left(g\left(\tfrac{1}{2}m_1 + m_2\right)l_1,\ -\tfrac{1}{2}gm_2l_2\right)\sin(z_k) + J_x V \dot{z}_k^2 + u_k\,,$$

where g is gravitational acceleration, $J_x \triangleq \frac{1}{2}m_2l_1l_2$, and V is the nonlinear velocity-coupling matrix:

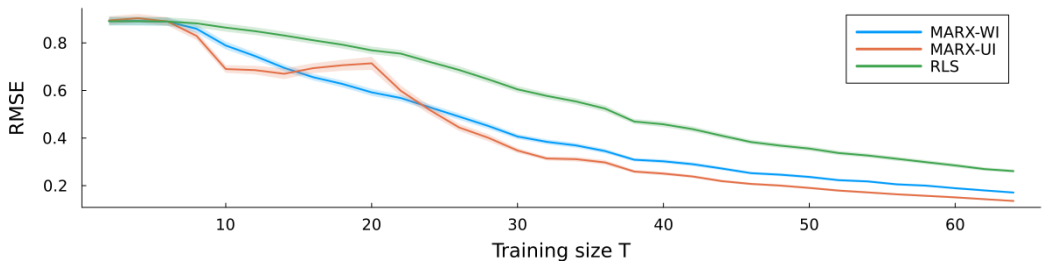
$$V = \begin{bmatrix} 0 & -\sin(z_{k,1} - z_{k,2}) \\ \sin(z_{k,1} - z_{k,2}) & 0 \end{bmatrix}.$$

6.3.3. Results

As in the verification experiment, Figure 12 shows the simulation errors for MARX-UI, MARX-WI, and RLS for both the double mass-spring-damper system Figure 12a) and the double pendulum system (Figure 12b). Convergence to stable performance is slower in both systems compared to the verification case. Nevertheless, both MARX variants outperform RLS and converge to similar levels of predictive performance. This confirms that the MARX model generalizes to more complex dynamical systems. As expected, the overall RMSE is higher for the nonlinear double pendulum system. A peak of performance loss is present for MARX-UI, which is more pronounced in the double mass-spring-damper system.



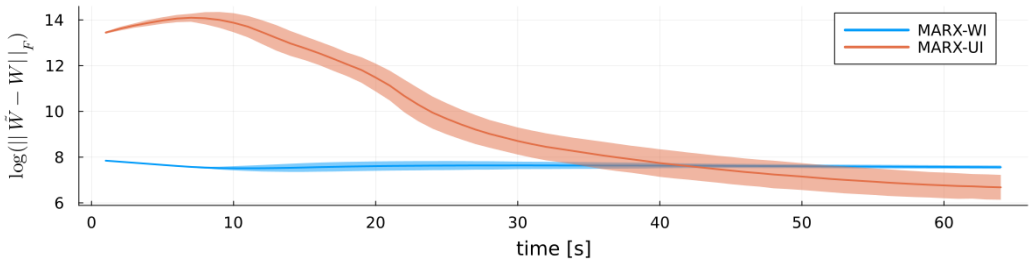
(**a**) Double mass-spring-damper system



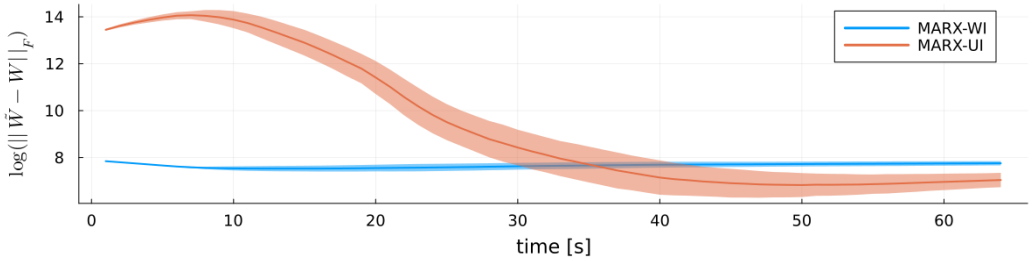
(**b**) Double pendulum system

Figure 12. Simulation errors (average RMSE) of all three estimators for each validation system, with ribbons indicating standard errors.

Figure 13 shows $\log(||\tilde{W} - W||_F)$ for both MARX-WI and MARX-UI for the validation systems. Initially, MARX-WI achieves better accuracy and lower variability than MARX-UI. Unlike in the verification setting, MARX-UI improves significantly over time and ultimately approaches similar estimation quality.



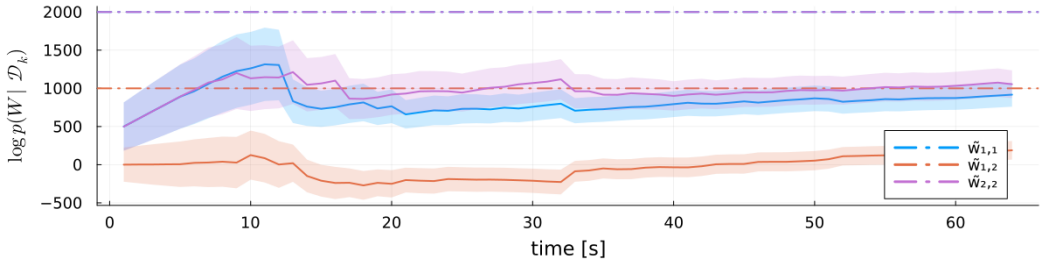
(**a**) Double mass-spring-damper system



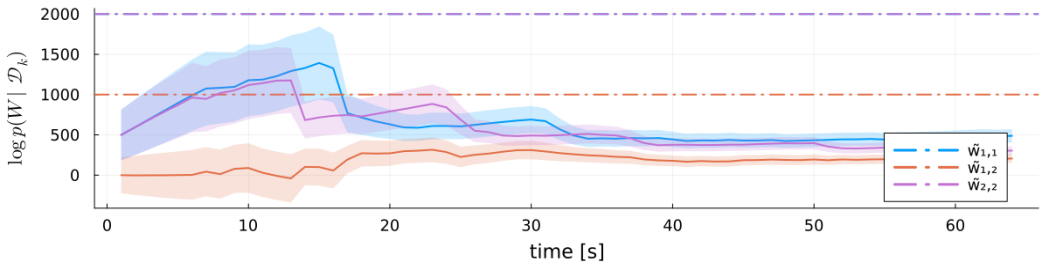
(**b**) Double pendulum system

Figure 13. Log-scale Frobenius norm of the error between the true coefficient matrix $\tilde{W}$ and its estimates W from each MARX estimator for each validation system. Ribbons represent standard errors.

Figure 14 illustrates estimates of $\tilde{W}$ by MARX-WI for a single Monte Carlo experiment ($T_{\text{train}} = 2^6$) for both systems. The model struggles with learning and initially shows high uncertainty, followed by a sharp reduction as learning progresses. This reflects the challenge of inferring observation noise structure in nonlinear systems from limited data.



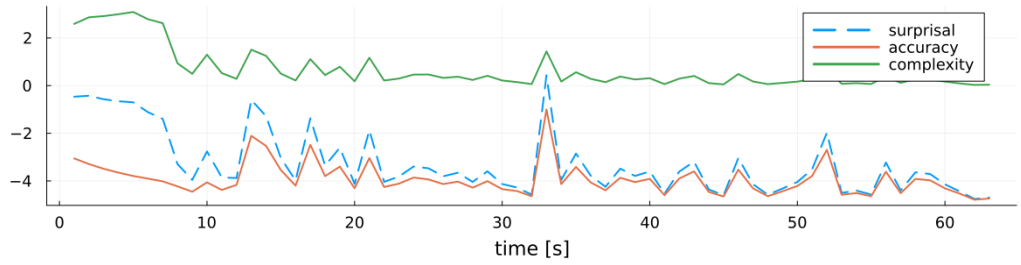
(**a**) Double mass-spring-damper system



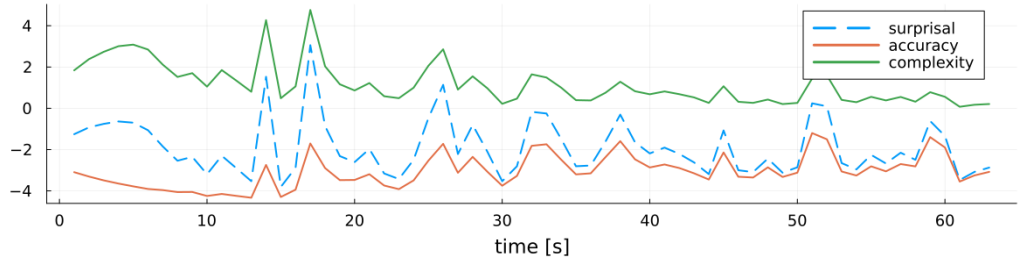
(**b**) Double pendulum system

Figure 14. Time series of $\tilde{W}$ estimates from MARX-WI for each validation system, with ribbons representing one standard deviation. Horizontal lines mark true parameter values.

Figure 15 displays the evolution of MARX-WI's surprisal and its decomposition into accuracy and complexity. The early learning phases show that surprisal reduction is dominated by decreasing model complexity. This trend is more difficult to sustain in the nonlinear system, where complexity remains elevated for longer. Later in training, fluctuations in surprisal are primarily driven by changes in accuracy.



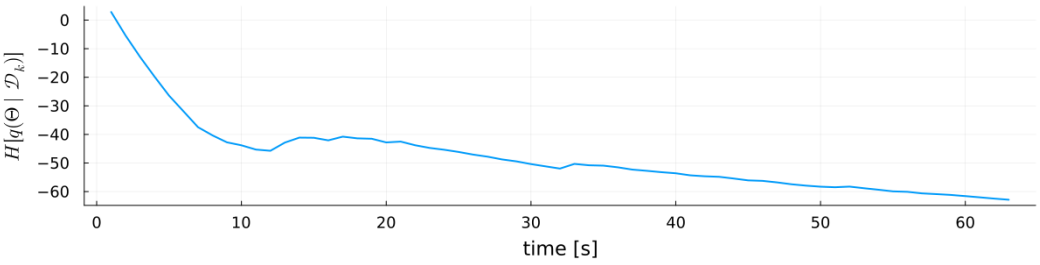
(**a**) Double mass-spring-damper system



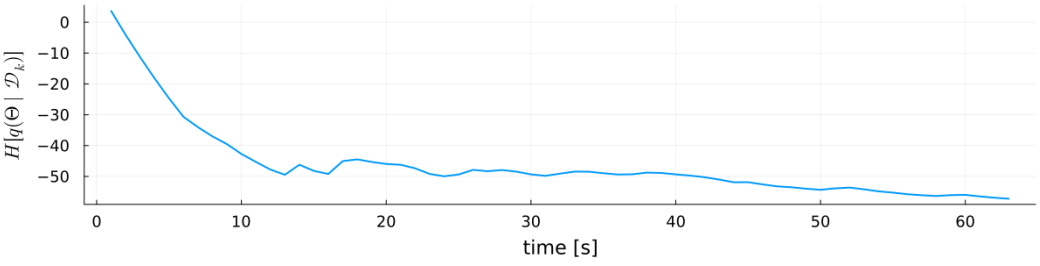
(**b**) Double pendulum system

Figure 15. Surprisal (dashed blue) and its decomposition into accuracy (red) and complexity (green) for MARX-WI over time for each validation system.

Finally, Figure 16 shows the entropy of the variational posterior $q(\Theta \mid \mathcal{D}_k)$ for each validation system. In both systems, MARX-WI rapidly reduces entropy, indicating fast convergence to informative parameter regions despite the different complexities of the systems.



(**a**) Double mass-spring-damper system



(**b**) Double pendulum system

Figure 16. Entropy of the MARX-WI model parameters over time for each validation system.

7. Discussion

The modular nature of the factor graph methodology provides substantial practical advantages. As demonstrated by Loeliger et al. [15], factor graphs facilitate the visual construction of complex algorithms by incorporating, eliminating, or merging established computational units. For example, the MARX model's factor graph (Figure 1) could be extended to support time-varying parameters by introducing state transition factor nodes

between the equality nodes over the parameters [24]. In multi-agent robotics, where sensors and actuators are spread across various platforms, each agent can update its local beliefs through message passing and share only the most informative summaries [44]. This targeted communication reduces bandwidth demands while enabling swift convergence to an accurate global model. Recent research highlights the importance of transmitting informative variational beliefs in multi-agent environments [22,45], facilitating scalable cooperative learning among heterogeneous agents. The resulting computational decentralization opens promising opportunities for federated system identification and coordination in multi-robot systems, especially when subject to privacy or bandwidth constraints [46–48].

7.1. Computational Efficiency

The dominant computational cost in our inference algorithm arises from the matrix inversion of Λ (4), which scales as $O(D_x^3)$ in the worst case. We benchmarked the update rule computations on a Julia-based implementation running on an Apple Macbook M1, averaging over 1,000,000 runs. For a state dimension of $D_x = 10$, updating the parameters for a single time step took approximately 2 nanoseconds (excluding garbage collection). Further computational savings are possible by adopting an information filter parameterization, where Ξ_k (A3) is stored instead of M_k (3) [49]. This approach defers the matrix inversion until M_k is explicitly needed, offering an efficiency boost, particularly in high-dimensional or resource-constrained scenarios.

7.2. Limitations

Despite its efficiency and modularity, our method has several limitations. First, it does not support fully Bayesian k-step ahead predictions. Computing joint posterior predictives over a longer horizon is intractable under the current formulation and is challenging as it requires marginalization over a (deeply) nested set of autoregressive coefficients. Second, the model is built on a linear multivariate autoregressive likelihood, which—while computationally efficient—limits its expressiveness. In systems characterized by strong nonlinearities, this assumption can lead to underfitting and reduced predictive performance. Lastly, although we explored both uninformative and weakly informative priors, the model remains sensitive to prior settings, particularly in data-scarce settings or during the early stages of recursive estimation. In these scenarios, poor prior choices can significantly degrade both convergence speed and final performance.

7.3. Future Work

Future work may explore extending the MARX framework to accommodate time-varying parameters by inserting state-transition factors between the equality nodes—analogous to prior work on univariate autoregressive models [24]. Another extension is to utilize the posterior distributions over the parameters to formulate a mutual information-based cost function for input signal design [10].

8. Conclusions

We presented a recursive Bayesian estimation procedure for multivariate autoregressive models with exogenous inputs. The method produces matrix-variate posterior distributions over both the model coefficients and the noise precision, allowing uncertainty to be explicitly propagated into future output predictions. We also demonstrated how these uncertainty estimates enable the analysis of individual factor nodes and edges within the model, making it possible to assess their contributions to the overall model score and to identify potential outliers. The ability to track sources of uncertainty online and evaluate their impact on output predictions is especially valuable for applications such as Bayesian optimal experimental design or information-theoretic adaptive control.

Author Contributions: T.N.N. contributed to the derivations, simulations, experimental results, and writing. W.M.K. contributed to the conception, direction, derivations, software, and writing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Eindhoven Artificial Intelligence Systems Institute.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: All data in this work is synthetic. For details on it was simulated, see the accompanying repository at https://github.com/biaslab/MDPI2025-MARX (accessed on 8 March 2025).

Acknowledgments: The authors gratefully acknowledge the support from Albert Podusenko.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MARX	multivariate autoregressive models with exogenous inputs
MARX-UI	MARX model with uninformative prior
MARX-WI	MARX model with weakly informative prior
VFE	variational free energy
BFE	Bethe Free Energy
RLS	recursive least squares
RMSE	Root Mean Square Error
ODE	Ordinary Differential Equation
KL	Kullback–Leibler

Appendix A. Parameter Estimation

Proof. The functional form of the likelihood is

$$p(y_k \mid \Theta, x_k) \propto \sqrt{|W|} \exp\left(-\frac{1}{2}\mathrm{tr}\left[W L_k\right]\right),$$

where $L_k \triangleq (y_k - A^\intercal x_k)(y_k - A^\intercal x_k)^\intercal$. The prior is

$$p(\Theta \mid \mathcal{D}_{k-1}) \propto \sqrt{|W|^{\nu_{t-1}+\bar{D}}} \exp\left(-\frac{1}{2}\mathrm{tr}\left[W\left(H_{k-1} + \Omega_{k-1}\right)\right]\right),$$

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

$$\begin{aligned} p(\Theta \mid \mathcal{D}_k) &\propto p(y_k \mid \Theta, x_k)\; p(\Theta \mid \mathcal{D}_{k-1}) \\ &\propto \sqrt{|W|^{\nu_{k-1}+1+\bar{D}}} \exp\left(-\frac{1}{2}\mathrm{tr}\left[W\left(L_k + H_{k-1} + \Omega_{k-1}\right)\right]\right). \end{aligned} \tag{A1}$$

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

$$\begin{aligned} L_k + H_{k-1} &= y_k y_k^\intercal - y_k x_k^\intercal A - A^\intercal x_k y_k^\intercal + A^\intercal x_k x_k^\intercal A \\ &\quad + A^\intercal \Lambda_{k-1} A - A^\intercal \Lambda_{k-1} M_{k-1} - M_{k-1}^\intercal \Lambda_{k-1} A + M_{k-1}^\intercal \Lambda_{k-1} M_{k-1} \\ &= A^\intercal(\Lambda_{k-1} + x_k x_k^\intercal) A - A^\intercal(x_k y_k^\intercal + \Lambda_{k-1} M_{k-1}) \\ &\quad - (M_{k-1}^\intercal \Lambda_{k-1} + y_k x_k^\intercal) A + y_k y_k^\intercal + M_{k-1}^\intercal \Lambda_{k-1} M_{k-1}\,. \end{aligned} \tag{A2}$$

Let $\Lambda_k \triangleq \Lambda_{k-1} + x_k x_k^\intercal$, $\Xi_k \triangleq x_k y_k^\intercal + \Lambda_{k-1} M_{k-1}$ and $M_k \triangleq \Lambda_k^{-1} \Xi_k$. Adding and subtracting $\Xi_k^\intercal \Lambda_k^{-1} \Xi_k$ to (A2) yields

$$\begin{aligned} L_k + H_{k-1} &= A^\intercal \Lambda_k A - A^\intercal \Xi_k - \Xi_k^\intercal A + \Xi_k^\intercal \Lambda_k^{-1} \Xi_k \\ &\quad - \Xi_k^\intercal \Lambda_k^{-1} \Xi_k + y_k y_k^\intercal + M_{k-1}^\intercal \Lambda_{k-1} M_{k-1} \\ &= (A - \Lambda_k^{-1} \Xi_k)^\intercal \Lambda_k (A - \Lambda_k^{-1} \Xi_k) \\ &\quad - M_k^\intercal \Lambda_k M_k + y_k y_k^\intercal + M_{k\text{-}1}^\intercal \Lambda_{k-1} M_{k-1} \,. \end{aligned} \tag{A3}$$

Plugging the above into (A1), we recognize the functional form of the matrix normal Wishart distribution:

$$\begin{aligned} &\sqrt{|W|^{\nu_k + \bar{D}}} \exp\big(-\frac{1}{2} \mathrm{tr}\big[W((A - M_k)^\intercal \Lambda_k (A - M_k) + \Omega_k) \big] \big) \\ &\qquad \propto \mathcal{MNW}(A, W \mid M_k, \Lambda_k^{-1}, \Omega_k^{-1}, \nu_k) \,, \end{aligned}$$

which parameters are

$$\begin{aligned} \nu_k &= \nu_{k-1} + 1 \,, \\ \Lambda_k &= \Lambda_{k-1} + x_k x_k^\intercal \,, \\ M_k &= (\Lambda_{k-1} + x_k x_k^\intercal)^{-1} (\Lambda_{k-1} M_{k-1} + x_k y_k^\intercal) \,, \text{ and} \\ \Omega_k &= \Omega_{k-1} + y_k y_k^\intercal + M_{k-1}^\intercal \Lambda_{k-1} M_{k-1} - M_k^\intercal \Lambda_k M_k \,. \end{aligned}$$

This concludes the proof. □

Appendix B. Backwards Message from Likelihood

Proof. The MARX likelihood function is

$$p(y_k \mid \Theta, x_k) \propto \sqrt{|W|} \exp\big(-\frac{1}{2} \mathrm{tr}\big[W L_k \big] \big) \,, \tag{A4}$$

where the completed square is

$$L_k \triangleq (y_k - A^\intercal x_k)(y_k - A^\intercal x_k)^\intercal = y_k y_k^\intercal - A^\intercal x_k y_k^\intercal - y_k x_k^\intercal A + A^\intercal x_k x_k^\intercal A \,.$$

Let $\bar{\Lambda}_k \triangleq x_k x_k^\intercal$, $\bar{\Xi}_k \triangleq x_k y_k^\intercal$ and $\bar{M}_k = \bar{\Lambda}_k^{-1} \bar{\Xi}_k$. Then adding and subtracting $\bar{\Xi}_k^\intercal \bar{\Lambda}_k \bar{\Xi}_k$ allows us to rewrite the square in terms of A:

$$L_k + \bar{\Xi}_k^\intercal \bar{\Lambda}_k^{-1} \bar{\Xi}_k - \bar{\Xi}_k^\intercal \bar{\Lambda}_k^{-1} \bar{\Xi}_k = y_k y_k^\intercal + (A - \bar{M}_k)^\intercal \bar{\Lambda}_k (A - \bar{M}_k) - \bar{\Xi}_k^\intercal \bar{\Lambda}_k^{-1} \bar{\Xi}_k \,.$$

The two remaining terms cancel:

$$\begin{aligned} y_k y_k^\intercal - \bar{\Xi}_k^\intercal \bar{\Lambda}_k^{-1} \bar{\Xi}_k &= y_k y_k^\intercal - y_k x_k^\intercal (x_k x_k^\intercal)^{-1} x_k y_k^\intercal \\ &= y_k y_k^\intercal - y_k I y_k^\intercal \\ &= 0_{D_y \times D_y} \,. \end{aligned}$$

If we define $\bar{\nu}_k \triangleq 1 - \bar{D}$ for $\bar{D} = D_x + D_y + 1$ and $\bar{\Omega}_k \triangleq 0_{D_y \times D_y}$, then we may recognize the functional form of a matrix normal Wishart in (A4):

$$\begin{aligned} p(y_k \mid \Theta, x_k) &\propto \sqrt{|W|^{\bar{\nu}_k + \bar{D}}} \exp\big(-\frac{1}{2} \mathrm{tr}\big[W((A - \bar{M}_k)^\intercal \bar{\Lambda}_k (A - \bar{M}_k) + \bar{\Omega}_k) \big] \big) \\ &\propto \mathcal{MNW}(A, W \mid \bar{M}_k, \bar{\Lambda}_k^{-1}, \bar{\Omega}_k^{-1}, \bar{\nu}_k) \,. \end{aligned}$$

This concludes the proof. □

Appendix C. Product of Matrix Normal Wishart Distributions

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

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

Their product is proportional to

$$\begin{aligned} p_1(\Theta)p_2(\Theta) &\propto \sqrt{|W|^{\nu_1+\bar{D}}}\exp\big(-\frac{1}{2}\mathrm{tr}\big[WL_1\big]\big)\sqrt{|W|^{\nu_2+\bar{D}}}\exp\big(-\frac{1}{2}\mathrm{tr}\big[WL_2\big]\big) \\ &= \sqrt{|W|^{\nu_3+\bar{D}}}\exp\big(-\frac{1}{2}\mathrm{tr}\big[W(L_1+L_2)\big]\big) \end{aligned}$$

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

$$\begin{aligned} L_1 + L_2 = {} & A^\intercal(\Lambda_1 + \Lambda_2)A - A^\intercal(\Lambda_1 M_1 + \Lambda_2 M_2) - (M_1^\intercal \Lambda_1 + M_2^\intercal \Lambda_2)A \\ & + M_1^\intercal \Lambda_1 M_1 + M_2^\intercal \Lambda_2 M_2 + \Omega_1 + \Omega_2\,. \end{aligned} \tag{A5}$$

Let $\Lambda_3 \triangleq \Lambda_1 + \Lambda_2$ and $\Theta_3 \triangleq \Lambda_1 M_1 + \Lambda_2 M_2$. Then,

$$(A - \Lambda_3^{-1}\Theta_3)^\intercal \Lambda_3 (A - \Lambda_3^{-1}\Theta_3) = A^\intercal \Lambda_3 A - A^\intercal \Theta_3 - \Theta_3^\intercal A + \Theta_3^\intercal \Lambda_3^{-1}\Theta_3\,.$$

Using $M_3 \triangleq \Lambda_3^{-1}\Theta_3$, (A5) can be written as

$$\begin{aligned} p_1(\Theta)p_2(\Theta) \propto \sqrt{|W|^{\nu_3+\bar{D}}}\exp\Big(-\frac{1}{2}\mathrm{tr}\big[W\big((A - M_3)^\intercal \Lambda_3 (A - M_3) \\ - \Theta_3^\intercal \Lambda_3^{-1}\Theta_3 + M_1^\intercal \Lambda_1 M_1 + M_2^\intercal \Lambda_2 M_2 + \Omega_1 + \Omega_2\big)\big]\Big)\,. \end{aligned} \tag{A6}$$

Note that $\Theta_3^\intercal \Lambda_3^{-1}\Theta_3 = \Theta_3^\intercal \Lambda_3^{-1}\Lambda_3\Lambda_3^{-1}\Theta_3 = M_3^\intercal \Lambda_3 M_3$. Let

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

Then (A6) may be recognized as an unnormalized matrix normal Wishart:

$$\begin{aligned} \sqrt{|W|^{\nu_3+\bar{D}}}\exp\big(-\frac{1}{2}\mathrm{tr}\big[W\big((A - M_3)^\intercal \Lambda_3 (A - M_3) + \Omega_3\big)\big]\big) \\ \propto \mathcal{MNW}\big(A, W \mid M_3, \Lambda_3^{-1}, \Omega_3^{-1}, \nu_3\big)\,. \end{aligned} \tag{A7}$$

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

Appendix D. Marginalization over A

Proof. The marginalization over A is

$$\begin{aligned} p(y_t \mid W, u_t, \mathcal{D}_k) &= \int p(y_t \mid \Theta, x_t) p(A \mid W, \mathcal{D}_k) \mathrm{d}A \\ &= \int \mathcal{N}\big(y_t \mid A^\intercal x_t, W^{-1}\big) \mathcal{MN}\big(A \mid M_k, \Lambda_k^{-1}, W^{-1}\big) \mathrm{d}A \\ &= \sqrt{(2\pi)^{-D_y(1+D_x)} |W|^{D_x+1} |\Lambda_k|^{D_y}} \\ &\quad \int \exp\Big(-\frac{1}{2}\mathrm{tr}\big[W(L_t + H_k)\big]\Big)\, \mathrm{d}A\,. \end{aligned} \tag{A8}$$

where the terms inside the trace are

$$\begin{aligned} L_t &\triangleq (y_t - A^\intercal x_t)(y_t - A^\intercal x_t)^\intercal \\ H_k &\triangleq (A - M_k)^\intercal \Lambda_k (A - M_k)\,. \end{aligned}$$

Expanding L_t and H_k and adding them yields

$$\begin{aligned} L_t + H_k &= y_t y_t^\intercal - A^\intercal x_t y_t^\intercal - y_t x_t^\intercal A + A^\intercal x_t x_t^\intercal A \\ &\quad + A^\intercal \Lambda_k A - A^\intercal \Lambda_k M_k - M_k^\intercal \Lambda_k A + M_k^\intercal \Lambda_k M_k \\ &= y_t y_t^\intercal + M_k^\intercal \Lambda_k M_k + A^\intercal (\Lambda_k + x_t x_t^\intercal) A \\ &\quad - A^\intercal (\Lambda_k M_k + x_t y_t^\intercal) - (\Lambda_k M_k + x_t y_t^\intercal)^\intercal A\,. \end{aligned}$$

Let $\Lambda_t \triangleq \Lambda_k + x_t x_t^\intercal$, $\Theta_t \triangleq \Lambda_k M_k + x_t y_t^\intercal$ and $M_t \triangleq \Lambda_t^{-1}\Theta_t$. Completing the square gives

$$L_t + H_k = (A - M_t)^\intercal \Lambda_t (A - M_t) - M_t^\intercal \Lambda_t M_t + y_t y_t^\intercal + M_k^\intercal \Lambda_k M_k\,.$$

Plugging this result into the integral in (A8) gives

$$\begin{aligned} \int \exp\Big(-\frac{1}{2}\mathrm{tr}\big[W(L_t + H_k)\big]\Big)\, \mathrm{d}A &= \exp\Big(-\frac{1}{2}\mathrm{tr}\big[W(y_t y_t^\intercal + M_k^\intercal \Lambda_k M_k - M_t^\intercal \Lambda_t M_t)\big]\Big) \\ &\quad \int \exp\Big(-\frac{1}{2}\mathrm{tr}\big[W(A - M_t)^\intercal \Lambda_t (A - M_t)\big]\Big) \mathrm{d}A\,. \end{aligned}$$

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

$$\int \exp\bigg(-\frac{1}{2}\mathrm{tr}\big[W(A - M_t)^\intercal \Lambda_t (A - M_t)\big]\bigg)\, \mathrm{d}A = \sqrt{\frac{(2\pi)^{D_y D_x}}{|W|^{D_x} |\Lambda_t|^{D_y}}}\,.$$

Using this result, the marginalization over A is

$$\begin{aligned} &\int p(y_t \mid \Theta, x_t) p(A \mid W, \mathcal{D}_k)\, \mathrm{d}A \\ &\quad = \sqrt{\frac{|W|}{(2\pi)^{D_y}}} \sqrt{|\Lambda_k|^{D_y} |\Lambda_t|^{-D_y}} \exp\Big(-\frac{1}{2}\mathrm{tr}\big[W(y_t y_t^\intercal + M_k^\intercal \Lambda_k M_k - M_t^\intercal \Lambda_t M_t)\big]\Big)\,. \end{aligned}$$

Note that, under the matrix determinant lemma,

$$|\Lambda_t| = |\Lambda_k + x_t x_t^\intercal| = |\Lambda_k|(1 + x_t^\intercal \Lambda_k^{-1} x_t)\,,$$

which implies that the product of determinants is

$$|\Lambda_k|^{D_y}|\Lambda_t|^{-D_y} = \left(1 + x_t^\intercal \Lambda_k^{-1} x_t\right)^{-D_y}.$$

Let $\lambda_t \triangleq (1 + x_t^\intercal \Lambda_k^{-1} x_t)^{-1}$. As W is D_y-dimensional, $|W|\lambda_t^{D_y} = |W\lambda_t|$. Furthermore, note that

$$\begin{aligned} M_t^\intercal \Lambda_t M_t = M_k^\intercal \Lambda_k (x_t x_t^\intercal + \Lambda_k)^{-1} \Lambda_k M_k + y_t x_t^\intercal (x_t x_t^\intercal + \Lambda_k)^{-1} \Lambda_k M_k \\ + M_k^\intercal \Lambda_k (x_t x_t^\intercal + \Lambda_k)^{-1} x_t y_t^\intercal + y_t x_t^\intercal (x_t x_t^\intercal + \Lambda_k)^{-1} x_t y_t^\intercal. \end{aligned}$$

Combining this with the other terms in the trace gives

$$\begin{aligned} & y_t y_t^\intercal + M_k^\intercal \Lambda_k M_k - M_t^\intercal \Lambda_t M_t \\ & \quad = M_k^\intercal \Lambda_k \left(I - (x_t x_t^\intercal + \Lambda_k)^{-1} \Lambda_k\right) M_k - y_t x_t^\intercal (x_t x_t^\intercal + \Lambda_k)^{-1} \Lambda_k M_k \\ & \quad\quad - M_k^\intercal \Lambda_k (x_t x_t^\intercal + \Lambda_k)^{-1} x_t y_t^\intercal + y_t \left(1 - x_t^\intercal (x_t x_t^\intercal + \Lambda_k)^{-1} x_t\right) y_t^\intercal. \end{aligned}$$

Using the Sherman–Morrison formula, we have

$$\begin{aligned} \left(1 - x_t^\intercal (x_t x_t^\intercal + \Lambda_k)^{-1} x_t\right) &= \left(1 - x_t^\intercal (\Lambda_k^{-1} - \frac{\Lambda_k^{-1} x_t x_t^\intercal \Lambda_k^{-1}}{1 + x_t^\intercal \Lambda_k^{-1} x_t}) x_t\right) \\ &= \left(1 + x_t^\intercal \Lambda_k^{-1} x_t\right)^{-1} \\ &= \lambda_t. \end{aligned}$$

Another application of Sherman–Morrison yields

$$\begin{aligned} I - (x_t x_t^\intercal + \Lambda_k)^{-1} \Lambda_k &= \left(\Lambda_k^{-1} - (\Lambda_k^{-1} - \frac{\Lambda_k^{-1} x_t x_t^\intercal \Lambda_k^{-1}}{1 + x_t^\intercal \Lambda_k^{-1} x_t})\right) \Lambda_k \\ &= \lambda_t \Lambda_k^{-1} x_t x_t^\intercal \Lambda_k^{-1}. \end{aligned}$$

A third Sherman–Morrison gives

$$\begin{aligned} \Lambda_k (x_t x_t^\intercal + \Lambda_k)^{-1} x_t &= \Lambda_k \left(\Lambda_k^{-1} - \frac{\Lambda_k^{-1} x_t x_t^\intercal \Lambda_k^{-1}}{1 + x_t^\intercal \Lambda_k^{-1} x_t}\right) x_t \\ &= I x_t - x_t \frac{x_t^\intercal \Lambda_k^{-1} x_t}{1 + x_t^\intercal \Lambda_k^{-1} x_t} \\ &= x_t \lambda_t. \end{aligned}$$

Using these three simplifications, we have

$$\begin{aligned} & \operatorname{tr}\left[W\left(y_t y_t^\intercal + M_k^\intercal \Lambda_k M_k - M_t^\intercal \Lambda_t M_t\right)\right] \\ & \quad = y_t^\intercal W \lambda_t y_t - y_t^\intercal W \lambda_t M_k^\intercal x_t - x_t^\intercal M_k W \lambda_t y_t + x_t^\intercal M_k W \lambda_t M_k^\intercal x_t \\ & \quad = \left(y_t - M_k^\intercal x_t\right)^\intercal W \lambda_t \left(y_t - M_k^\intercal x_t\right). \end{aligned} \tag{A9}$$

Plugging (A9) into (A8) yields

$$\begin{aligned} p(y_t \mid W, u_t, \mathcal{D}_k) &= \sqrt{\frac{|W\lambda_t|}{(2\pi)^{D_y}}} \exp\left(-\frac{\lambda_t}{2}(y_t - M_k^\intercal x_t)^\intercal W (y_t - M_k^\intercal x_t)\right) \\ &= \mathcal{N}\left(y_t \mid M_k^\intercal x_t, (W\lambda_t)^{-1}\right). \end{aligned}$$

This concludes the proof. □

Appendix E. Marginalization over W

Proof. The marginalization over W is

$$\begin{aligned}
&\int \mathcal{N}(y_t \mid M_k^\intercal x_t, (W\lambda_t)^{-1})\mathcal{W}(W \mid \Omega_k^{-1}, \nu_k)\mathrm{d}W \\
&\quad= \int \sqrt{\frac{|W\lambda_t|}{(2\pi)^{D_y}}} \exp\big(-\frac{\lambda_t}{2}(y_t - M_k^\intercal x_t)^\intercal W(y_t - M_k^\intercal x_t)\big) \\
&\quad\quad \frac{1}{\Gamma_{D_y}(\frac{\nu_k}{2})}\sqrt{\frac{|\Omega_k|^{\nu_k}|W|^{\nu_k - D_y - 1}}{2^{\nu_k D_y}}} \exp\big(-\frac{1}{2}\mathrm{tr}\big[W\Omega_k\big]\big)\mathrm{d}W \\
&\quad= \frac{1}{\Gamma_{D_y}(\frac{\nu_k}{2})}\sqrt{\frac{|\Omega_k|^{\nu_k}\lambda_t^{D_y}}{(2\pi)^{D_y}2^{\nu_k D_y}}} \int \sqrt{|W|^{\nu_k + 1 - D_y - 1}} \\
&\quad\quad \exp\big(-\frac{1}{2}\mathrm{tr}\big[W(\Omega_k + \lambda_t(y_t - M_k^\intercal x_t)(y_t - M_k^\intercal x_t)^\intercal)\big]\big)\mathrm{d}W \\
&\quad= \frac{1}{\Gamma_{D_y}(\frac{\nu_k}{2})}\sqrt{\frac{|\Omega_k|^{\nu_k}\lambda_t^{D_y}}{(2\pi)^{D_y}2^{\nu_k D_y}}}\Gamma_{D_y}(\frac{\nu_k + 1}{2})\sqrt{2^{(\nu_k+1)D_y}} \\
&\quad\quad \sqrt{\big|\Omega_k + \lambda_t(y_t - M_k^\intercal x_t)(y_t - M_k^\intercal x_t)^\intercal\big|^{-(\nu_k+1)}},
\end{aligned} \tag{A10}$$

where we made use of the normalization factor of a Wishart distribution. Note that

$$\sqrt{\frac{2^{(\nu_k+1)D_y}}{(2\pi)^{D_y}2^{\nu_k D_y}}} = \sqrt{\frac{2^{D_y}}{2^{D_y}\pi^{D_y}}} = \sqrt{\frac{1}{\pi^{D_y}}}. \tag{A11}$$

Let $\eta_t \triangleq \nu_k - D_y + 1$. Then,

$$\frac{\Gamma_{D_y}(\frac{\nu_k+1}{2})}{\Gamma_{D_y}(\frac{\nu_k}{2})} = \frac{\Gamma_{D_y}(\frac{\eta_t+D_y}{2})}{\Gamma_{D_y}(\frac{\eta_t+D_y-1}{2})}. \tag{A12}$$

The determinants simplify as follows:

$$\begin{aligned}
&\sqrt{|\Omega_k|^{\nu_k}|\Omega_k + \lambda_t(y_t - M_k^\intercal x_t)(y_t - M_k^\intercal x_t)^\intercal|^{-(\nu_k+1)}} \\
&\quad= \sqrt{|\lambda_t(y_t - M_k^\intercal x_t)(y_t - M_k^\intercal x_t)^\intercal\Omega_k^{-1} + I|^{-(\nu_k+1)}|\Omega_k^{-1}|},
\end{aligned} \tag{A13}$$

and then, using the matrix determinant lemma, we have

$$\begin{aligned}
&|\lambda_t(y_t - M_k^\intercal x_t)(y_t - M_k^\intercal x_t)^\intercal\Omega_k^{-1} + I|^{-(\nu_k+1)} \\
&\quad= \big((y_t - M_k^\intercal x_t)^\intercal\Omega_k^{-1}\lambda_t(y_t - M_k^\intercal x_t) + 1\big)^{-(\eta_t+D_y)/2}.
\end{aligned} \tag{A14}$$

With Equations (A11)–(A14), we may write (A10) as

$$\begin{aligned}
&\int \mathcal{N}\big(y_t \mid M_k x_t, (W\lambda_t)^{-1}\big)\mathcal{W}(W \mid \Omega_k^{-1}, \nu_k)\mathrm{d}W \\
&= \sqrt{\frac{|\Omega_k^{-1}|}{\pi^{D_y}}}\frac{\Gamma_{D_y}(\frac{\eta_t+D_y}{2})}{\Gamma_{D_y}(\frac{\eta_t+D_y-1}{2})}\sqrt{\lambda_t^{D_y}} \\
&\quad \big(1+(y_t-M_k^\intercal x_t)^\intercal \Omega_k^{-1}\lambda_t(y_t-M_k^\intercal x_t)\big)^{-(\eta_t+D_y)/2} \\
&= \sqrt{\frac{|\eta_t\Omega_k^{-1}\lambda_t|}{(\eta_t\pi)^{D_y}}}\frac{\Gamma_{D_y}((\eta_t+D_y)/2)}{\Gamma_{D_y}((\eta_t+D_y-1)/2)} \\
&\quad \big(1+\frac{1}{\eta_t}(y_t-M_k^\intercal x_t)^\intercal \eta_t\Omega_k^{-1}\lambda_t(y_t-M_k^\intercal x_t)\big)^{-(\eta_t+D_y)/2} \\
&= \mathcal{T}(y_t \mid \mu_t, \Psi_t^{-1}, \eta_t)\,,
\end{aligned}$$

where $\mu_t \triangleq M_k^\intercal x_t$, $\Psi_t \triangleq \eta_t\Omega_k^{-1}\lambda_t$. □

Appendix F. Cross-Entropy of a Matrix Normal Wishart Relative to a Matrix Normal Wishart

Proof. The functional form of a matrix normal Wishart with general parameters M, Λ, Ω, and ν is

$$\begin{aligned}
p(\Theta) &= \mathcal{MNW}(A, W \mid M, \Lambda^{-1}, \Omega^{-1}, \nu) \\
&= \sqrt{\frac{|\Lambda|^{D_y}|\Omega|^{\nu}|W|^{\nu+D_x-D_y-1}}{2^{(\nu+D_x)D_y}\pi^{D_xD_y}}}\frac{1}{\Gamma_{D_y}(\frac{\nu}{2})}\exp\big(-\frac{1}{2}\mathrm{tr}\big[W\big((A-M)^\intercal\Lambda(A-M)+\Omega\big)\big]\big)\,.
\end{aligned}$$

Consider two matrix normal Wishart distributions over the same parameters Θ:

$$\begin{aligned}
q(\Theta) &= \mathcal{MNW}(A, W \mid M_q, \Lambda_q^{-1}, \Omega_q^{-1}, \nu_q) \\
p(\Theta) &= \mathcal{MNW}(A, W \mid M_p, \Lambda_p^{-1}, \Omega_p^{-1}, \nu_p)\,.
\end{aligned}$$

The differential cross-entropy $H[q,p]$ of q relative to p is

$$\begin{aligned}
H[q,p] &= \mathbb{E}_{q(\Theta)}\Big[-\log p(\Theta)\Big] \\
&= -\frac{1}{2}D_y\log|\Lambda_p| - \frac{1}{2}\nu_p\log|\Omega_p| - \frac{1}{2}(\nu_p+D_x-D_y-1)\mathbb{E}_q\Big[\log|W|\Big] \\
&\quad + \frac{1}{2}(\nu_p+D_x)D_y\log 2 + \frac{1}{2}D_xD_y\log\pi + \log\Gamma_{D_y}(\frac{\nu_p}{2}) \\
&\quad + \frac{1}{2}\mathbb{E}_q\Big[\mathrm{tr}\big[W\big((A-M_p)^\intercal\Lambda_p(A-M_p)+\Omega_p\big)\big]\Big]\,.
\end{aligned} \tag{A15}$$

The first expectation is the expectation of a Wishart log-determinant [28]:

$$\mathbb{E}_{q(\Theta)}\Big[\log|W|\Big] = \mathbb{E}_{q(W)}\Big[\mathbb{E}_{q(A\mid W)}\Big[\log|W|\Big]\Big] = \psi_{D_y}(\frac{\nu_q}{2}) + D_y\log 2 - \log|\Omega_q|\,, \tag{A16}$$

where ψ_{D_y} is the multivariate digamma function of dimension D_y. For the second expectation, we first define the following expectations:

$$\mathbb{E}_{q(A \mid W)}\Big[\, A \,\Big] = M_q \tag{A17}$$

$$\mathbb{E}_{q(A \mid W)}\Big[\, A^\intercal \,\Big] = M_q^\intercal \tag{A18}$$

$$\mathbb{E}_{q(A \mid W)}\Big[\, A^\intercal B A \,\Big] = M_q^\intercal B M_q + \mathrm{tr}(\Lambda_q^{-1} B^\intercal) W^{-1} \tag{A19}$$

$$\mathbb{E}_{q(W)}\Big[\, W \,\Big] = \nu_q \Omega_q^{-1}\,, \tag{A20}$$

for appropriately dimensioned matrix B. Equation (A19) is a property of matrix normal distributions [28]. We apply $\mathbb{E}_q[\,\mathrm{tr}(\cdot)\,] = \mathrm{tr}(\mathbb{E}_q[\,\cdot\,])$ [27], and make use of the factorization of a matrix normal Wishart:

$$\begin{aligned}
&\mathbb{E}_{q(\Theta)}\Big[\,\mathrm{tr}\Big(W\big((A - M_p)^\intercal \Lambda_p (A - M_p) + \Omega_p\big)\Big)\,\Big] \\
&\quad = \mathrm{tr}\Big(\mathbb{E}_{q(\Theta)}\Big[\, W\big((A - M_p)^\intercal \Lambda_p (A - M_p) + \Omega_p\big)\,\Big]\Big) \\
&\quad = \mathrm{tr}\Big(\mathbb{E}_{q(W)}\Big[\, \mathbb{E}_{q(A \mid W)}\Big[\, W\big((A - M_p)^\intercal \Lambda_p (A - M_p) + \Omega_p\big)\,\Big]\,\Big]\Big) \\
&\quad = \mathrm{tr}\Big(\mathbb{E}_{q(W)}\Big[\, W\mathbb{E}_{q(A \mid W)}\Big[\, (A - M_p)^\intercal \Lambda_p (A - M_p)\,\Big] + W\Omega_p\,\Big]\Big)\,.
\end{aligned} \tag{A21}$$

We expand the term in the inner expectation and plug in Equations (A17)–(A19):

$$\begin{aligned}
&\mathbb{E}_{q(A \mid W)}\Big[\, (A - M_p)^\intercal \Lambda_p (A - M_p)\,\Big] \\
&\quad = \mathbb{E}_{q(A \mid W)}\Big[\, A^\intercal \Lambda_p A - A^\intercal \Lambda_p M_p - M_p^\intercal \Lambda_p A + M_p^\intercal \Lambda_p M_p\,\Big] \\
&\quad = M_q^\intercal \Lambda_p M_q + \mathrm{tr}(\Lambda_q^{-1} \Lambda_p^\intercal) W^{-1} - M_q^\intercal \Lambda_p M_p - M_p^\intercal \Lambda_p M_q + M_p^\intercal \Lambda_p M_p \\
&\quad = (M_q - M_p)^\intercal \Lambda_p (M_q - M_p) + \mathrm{tr}(\Lambda_q^{-1} \Lambda_p^\intercal) W^{-1}\,.
\end{aligned} \tag{A22}$$

We plug (A22) into (A21), expand, and use (A20) to resolve the remaining expectations:

$$\begin{aligned}
&\mathrm{tr}\Big(\mathbb{E}_{q(W)}\Big[\, W\mathbb{E}_{q(A \mid W)}\Big[\, (A - M_p)^\intercal \Lambda_p (A - M_p)\,\Big] + W\Omega_p\,\Big]\Big) \\
&\quad = \mathrm{tr}\Big(\mathbb{E}_{q(W)}\Big[\, W(M_q - M_p)^\intercal \Lambda_p (M_q - M_p) + \mathrm{tr}(\Lambda_q^{-1} \Lambda_p^\intercal) W W^{-1} + W\Omega_p\,\Big]\Big) \\
&\quad = \mathrm{tr}\Big(\nu_q \Omega_q^{-1} (M_q - M_p)^\intercal \Lambda_p (M_q - M_p) + \mathrm{tr}(\Lambda_q^{-1} \Lambda_p^\intercal) I_{D_y} + \nu_q \Omega_q^{-1} \Omega_p\Big) \\
&\quad = \nu_q \mathrm{tr}\Big(\Omega_q^{-1} (M_q - M_p)^\intercal \Lambda_p (M_q - M_p)\Big) + D_y \mathrm{tr}(\Lambda_q^{-1} \Lambda_p^\intercal) + \nu_q \mathrm{tr}(\Omega_q^{-1} \Omega_p)\,.
\end{aligned} \tag{A23}$$

We plug (A23) and (A16) into (A15) to yield the differential cross-entropy:

$$\begin{aligned}
H[q,p] &= -\frac{1}{2}D_y \log|\Lambda_p| - \frac{1}{2}\nu_p \log|\Omega_p| \\
&\quad - \frac{1}{2}(\nu_p + D_x - D_y - 1)\Big(\psi_{D_y}(\frac{\nu_q}{2}) + D_y \log 2 - \log|\Omega_q|\Big) \\
&\quad + \frac{1}{2}(\nu_p + D_x)D_y \log 2 + \frac{1}{2}D_x D_y \log \pi + \log \Gamma_{D_y}(\frac{\nu_p}{2}) \\
&\quad + \frac{1}{2}\nu_q \mathrm{tr}\Big(\Omega_q^{-1}(M_q - M_p)^\intercal \Lambda_p (M_q - M_p)\Big) + \frac{1}{2}D_y \mathrm{tr}(\Lambda_q^{-1}\Lambda_p^\intercal) \\
&\quad + \frac{1}{2}\nu_q \mathrm{tr}(\Omega_q^{-1}\Omega_p) \\
&= -\frac{1}{2}D_y \log|\Lambda_p| + \frac{1}{2}(\nu_p + D_x - D_y - 1)\log|\Omega_q| - \frac{1}{2}\nu_p \log|\Omega_p| \\
&\quad + \frac{1}{2}(D_y + 1)D_y \log 2 + \frac{1}{2}D_x D_y \log \pi \\
&\quad + \log \Gamma_{D_y}(\frac{\nu_p}{2}) - \frac{1}{2}(\nu_p + D_x - D_y - 1)\psi_{D_y}(\frac{\nu_q}{2}) \\
&\quad + \frac{1}{2}\nu_q \mathrm{tr}\Big(\Omega_q^{-1}(M_q - M_p)^\intercal \Lambda_p (M_q - M_p)\Big) \\
&\quad + \frac{1}{2}D_y \mathrm{tr}(\Lambda_q^{-1}\Lambda_p^\intercal) + \nu_q \mathrm{tr}(\Omega_q^{-1}\Omega_p)\,.
\end{aligned} \tag{A24}$$

This concludes the proof. □

Appendix G. Entropy of a Matrix Normal Wishart

Proof. Consider a matrix normal Wishart distribution:

$$q(\Theta) = \mathcal{MNW}(A, W \mid M_q, \Lambda_q^{-1}, \Omega_q^{-1}, \nu_q)\,.$$

By definition, a differential entropy $H[q]$ of a distribution q is a special case of a differential cross-entropy $H[q,p]$ of q from another distribution p, where $p = q$, i.e., $H[q] = H[q,q]$. Plugging in (the parameters of) $p = q$ into (A24) from Appendix F, we get the entropy:

$$\begin{aligned}
H[q] &= -\frac{1}{2}D_y \log|\Lambda_q| + \frac{1}{2}(\nu_q + D_x - D_y - 1)\log|\Omega_q| - \frac{1}{2}\nu_q \log|\Omega_q| \\
&\quad + \frac{1}{2}(D_y + 1)D_y \log 2 + \frac{1}{2}D_x D_y \log \pi \\
&\quad + \log \Gamma_{D_y}(\frac{\nu_q}{2}) - \frac{1}{2}(\nu_q + D_x - D_y - 1)\psi_{D_y}(\frac{\nu_q}{2}) \\
&\quad + \underbrace{\frac{1}{2}\nu_q \mathrm{tr}\Big(\Omega_q^{-1}(M_q - M_q)^\intercal \Lambda_q (M_q - M_q)\Big)}_{=0} + \frac{1}{2}D_y \underbrace{\mathrm{tr}(\Lambda_q^{-1}\Lambda_q^\intercal)}_{=D_x} + \frac{1}{2}\nu_q \underbrace{\mathrm{tr}(\Omega_q^{-1}\Omega_q)}_{=D_y} \\
&= -\frac{1}{2}D_y \log|\Lambda_q| + \frac{1}{2}(D_x - D_y - 1)\log|\Omega_q| \\
&\quad + \frac{1}{2}(D_y + 1)D_y \log 2 + \frac{1}{2}D_x D_y \log \pi + \frac{1}{2}(D_x + \nu_q)D_y \\
&\quad + \log \Gamma_{D_y}(\frac{\nu_q}{2}) - \frac{1}{2}(\nu_q + D_x - D_y - 1)\psi_{D_y}(\frac{\nu_q}{2})\,.
\end{aligned}$$

This concludes the proof. □

Appendix H. KL-Divergence of a Matrix Normal Wishart from a Matrix Normal Wishart

Proof. Consider two matrix normal Wishart distributions over the same parameters Θ:

$$\begin{aligned} q(\Theta) &= \mathcal{MNW}(A, W \mid M_q, \Lambda_q^{-1}, \Omega_q^{-1}, \nu_q) \\ p(\Theta) &= \mathcal{MNW}(A, W \mid M_p, \Lambda_p^{-1}, \Omega_p^{-1}, \nu_p)\,. \end{aligned}$$

By definition, a KL-divergence $D_{KL}[q||p]$ of a distribution q from another distribution p is the difference between the differential cross-entropy $H[q,p]$ of q from p (A24) and the entropy of q (14) [50]:

$$\begin{aligned} D_{KL}[q||p] = &-\frac{1}{2}D_y \log|\Lambda_p| + \frac{1}{2}(\nu_p + D_x - D_y - 1)\log|\Omega_q| - \frac{1}{2}\nu_p \log|\Omega_p| \\ &+ \frac{1}{2}(D_y+1)D_y \log 2 + \frac{1}{2}D_x D_y \log\pi \\ &+ \log\Gamma_{D_y}(\frac{\nu_p}{2}) - \frac{1}{2}(\nu_p + D_x - D_y - 1)\psi_{D_y}(\frac{\nu_q}{2}) \\ &+ \frac{1}{2}\nu_q \mathrm{tr}\Big(\Omega_q^{-1}(M_q - M_p)^\intercal \Lambda_p (M_q - M_p)\Big) + \frac{1}{2}D_y \mathrm{tr}(\Lambda_q^{-1}\Lambda_p^\intercal) + \frac{1}{2}\nu_q \mathrm{tr}(\Omega_q^{-1}\Omega_p) \\ &+ \frac{1}{2}D_y \log|\Lambda_q| - \frac{1}{2}(D_x - D_y - 1)\log|\Omega_q| \\ &- \frac{1}{2}(D_y+1)D_y \log 2 - \frac{1}{2}D_x D_y \log\pi - \frac{1}{2}(D_x + \nu_q)D_y \\ &- \log\Gamma_{D_y}(\frac{\nu_q}{2}) + \frac{1}{2}(\nu_q + D_x - D_y - 1)\psi_{D_y}(\frac{\nu_q}{2}) \\ = &\frac{1}{2}D_y \log\frac{|\Lambda_q|}{|\Lambda_p|} + \frac{1}{2}\nu_p \log\frac{|\Omega_q|}{|\Omega_p|} - \frac{1}{2}(D_x + \nu_q)D_y \\ &- \log\Gamma_{D_y}(\frac{\nu_q}{2}) + \log\Gamma_{D_y}(\frac{\nu_p}{2}) + \frac{1}{2}(\nu_q - \nu_p)\psi_{D_y}(\frac{\nu_q}{2}) \\ &+ \frac{1}{2}\nu_q \mathrm{tr}\Big(\Omega_q^{-1}(M_q - M_p)^\intercal \Lambda_p (M_q - M_p)\Big) + \frac{1}{2}D_y \mathrm{tr}(\Lambda_q^{-1}\Lambda_p^\intercal) + \nu_q \mathrm{tr}(\Omega_q^{-1}\Omega_p)\,. \end{aligned}$$

This concludes the proof. □

Appendix I. Cross-Entropy of a Matrix Normal Wishart Relative to a Multivariate Normal

Proof. Consider a matrix normal Wishart distribution q and a multivariate normal distribution p:

$$\begin{aligned} q(\Theta) &= \mathcal{MNW}(A, W \mid M_q, \Lambda_q^{-1}, \Omega_q^{-1}, \nu_q) \\ p(y \mid \Theta, x) &= \mathcal{N}(y \mid A^\intercal x, W^{-1})\,. \end{aligned}$$

The differential cross-entropy $H[q,p]$ of q relative to p is

$$\begin{aligned} H[q,p] &= \mathbb{E}_{q(\Theta)}\Big[-\log p(y \mid \Theta, x)\Big] \\ &= -\frac{1}{2}\mathbb{E}_{q(\Theta)}\Big[\log|W|\Big] + \frac{1}{2}D_y \log 2 + \frac{1}{2}D_y \log\pi \\ &\quad + \frac{1}{2}\mathbb{E}_{q(\Theta)}\Big[(y - A^\intercal x)^\intercal W (y - A^\intercal x)\Big]\,. \end{aligned} \tag{A25}$$

The first expectation again is the expectation of a Wishart log-determinant [28] (A16). For the second expectation, we make use of the factorization of a matrix normal Wishart, bring

the term (a scalar) in trace form, apply $\mathbb{E}_q[\,\mathrm{tr}(\cdot)\,] = \mathrm{tr}(\mathbb{E}_q[\,\cdot\,])$ [27] and the cyclic property of tr, and rearrange as follows:

$$\begin{aligned}
&\mathbb{E}_{q(\Theta)}\Big[\,(y - A^\intercal x)^\intercal W (y - A^\intercal x)\,\Big] \\
&\quad = \mathbb{E}_{q(W)}\Big[\,\mathbb{E}_{q(A\,|\,W)}\Big[\,(y - A^\intercal x)^\intercal W (y - A^\intercal x)\,\Big]\,\Big] \\
&\quad = \mathrm{tr}\Big(\mathbb{E}_{q(W)}\Big[\,\mathbb{E}_{q(A\,|\,W)}\Big[\,W (y - A^\intercal x)(y - A^\intercal x)^\intercal\,\Big]\,\Big]\Big) \\
&\quad = \mathrm{tr}\Big(\mathbb{E}_{q(W)}\Big[\,W\mathbb{E}_{q(A\,|\,W)}\Big[\,(y - A^\intercal x)(y - A^\intercal x)^\intercal\,\Big]\,\Big]\Big)\,. \qquad \text{(A26)}
\end{aligned}$$

We expand the term in the inner expectation and plug in (A17) and (A19) (with $B = xx^\intercal$):

$$\begin{aligned}
&\mathbb{E}_{q(A\,|\,W)}\Big[\,(y - A^\intercal x)(y - A^\intercal x)^\intercal\,\Big] \\
&\quad = yy^\intercal - \mathbb{E}_{q(A\,|\,W)}\Big[\,yx^\intercal A\,\Big] - \mathbb{E}_{q(A\,|\,W)}\Big[\,A^\intercal xy^\intercal\,\Big] + \mathbb{E}_{q(A\,|\,W)}\Big[\,A^\intercal xx^\intercal A\,\Big] \\
&\quad = yy^\intercal - yx^\intercal \mathbb{E}_{q(A\,|\,W)}\Big[\,A\,\Big] - \mathbb{E}_{q(A\,|\,W)}\Big[\,A^\intercal\,\Big]xy^\intercal + \mathbb{E}_{q(A\,|\,W)}\Big[\,A^\intercal xx^\intercal A\,\Big] \\
&\quad = yy^\intercal - yx^\intercal M_q - M_q^\intercal xy^\intercal + M_q^\intercal xx^\intercal M_q + \mathrm{tr}(\Lambda_q^{-1}(xx^\intercal)^\intercal)W^{-1} \\
&\quad = (y - M_q^\intercal x)(y - M_q^\intercal x)^\intercal + \mathrm{tr}(\Lambda_q^{-1}xx^\intercal)W^{-1} \\
&\quad = (y - M_q^\intercal x)(y - M_q^\intercal x)^\intercal + x^\intercal \Lambda_q^{-1} x W^{-1}\,. \qquad \text{(A27)}
\end{aligned}$$

Note that all terms are within a trace, so we can apply the cyclic property of the trace, and $x^\intercal x$ is a scalar. We plug in (A27) into (A26) and use (A20) to solve the expectation:

$$\begin{aligned}
&\mathbb{E}_{q(\Theta)}\Big[\,(y - A^\intercal x)^\intercal W (y - A^\intercal x)\,\Big] \\
&\quad = \mathrm{tr}\Big(\mathbb{E}_{q(W)}\Big[\,W((y - M_q^\intercal x)(y - M_q^\intercal x)^\intercal + x^\intercal \Lambda_q^{-1} x W^{-1})\,\Big]\Big) \\
&\quad = \mathrm{tr}\Big(\mathbb{E}_{q(W)}\Big[\,W(y - M_q^\intercal x)(y - M_q^\intercal x)^\intercal + x^\intercal \Lambda_q^{-1} x \underbrace{WW^{-1}}_{=I_{D_y}}\,\Big]\Big) \\
&\quad = \mathrm{tr}\Big(\mathbb{E}_{q(W)}\Big[\,W(y - M_q^\intercal x)(y - M_q^\intercal x)^\intercal\,\Big]\Big) + x^\intercal \Lambda_q^{-1} x \mathrm{tr}(I_{D_y}) \\
&\quad = \mathrm{tr}\Big(\nu_q \Omega_q^{-1}(y - M_q^\intercal x)(y - M_q^\intercal x)^\intercal\Big) + x^\intercal \Lambda_q^{-1} x D_y \\
&\quad = \nu_q (y - M_q^\intercal x)^\intercal \Omega_q^{-1}(y - M_q^\intercal x) + x^\intercal \Lambda_q^{-1} x D_y\,. \qquad \text{(A28)}
\end{aligned}$$

We plug (A28) and (A16) into (A25) to yield the differential cross-entropy:

$$\begin{aligned}
H[q,p] = &-\frac{1}{2}\psi_{D_y}\Big(\frac{\nu_q}{2}\Big) + \frac{1}{2}\log|\Omega_q| + \frac{1}{2}D_y \log \pi \\
&+ \frac{1}{2}\nu_q (y - M_q^\intercal x)^\intercal \Omega_q^{-1}(y - M_q^\intercal x) + \frac{1}{2}x^\intercal \Lambda_q^{-1} x D_y\,.
\end{aligned}$$

This concludes the proof. □

References

1. Nisslbeck, T.N.; Kouw, W.M. Online Bayesian system identification in multivariate autoregressive models via message passing. (*accepted*). In Proceedings of the European Control Conference, Thessaloniki, Greece, 24–27 June 2025; IEEE: New York, NY, USA, 2025.
2. Tiao, G.C.; Zellner, A. On the Bayesian estimation of multivariate regression. *J. R. Stat. Soc. Ser. B* **1964**, *26*, 277–285. [CrossRef]
3. Hannan, E.J.; McDougall, A.; Poskitt, D.S. Recursive estimation of autoregressions. *J. R. Stat. Soc. Ser. B* **1989**, *51*, 217–233. [CrossRef]
4. Karlsson, S. Forecasting with Bayesian vector autoregression. *Handb. Econ. Forecast.* **2013**, *2*, 791–897.

5. Nisslbeck, T.N.; Kouw, W.M. Coupled autoregressive active inference agents for control of multi-joint dynamical systems. In Proceedings of the International Workshop on Active Inference, Oxford, UK, 9–11 September 2024; Springer: Berlin/Heidelberg, Germany, 2024.
6. Barber, D. *Bayesian Reasoning and Machine Learning*; Cambridge University Press: Cambridge, UK, 2012.
7. Hecq, A.; Issler, J.V.; Telg, S. Mixed causal–noncausal autoregressions with exogenous regressors. *J. Appl. Econom.* **2020**, *35*, 328–343. [CrossRef]
8. Penny, W.; Harrison, L. Multivariate autoregressive models. In *Statistical Parametric Mapping: The Analysis of Functional Brain Images*; Academic Press: Amsterdam, The Netherlands, 2007; pp. 534–540.
9. Shaarawy, S.M.; Ali, S.S. Bayesian identification of multivariate autoregressive processes. *Commun. Stat. Methods* **2008**, *37*, 791–802. [CrossRef]
10. Chaloner, K.; Verdinelli, I. Bayesian experimental design: A review. *Stat. Sci.* **1995**, *10*, 273–304. [CrossRef]
11. Williams, G.; Drews, P.; Goldfain, B.; Rehg, J.M.; Theodorou, E.A. Information-theoretic model predictive control: Theory and applications to autonomous driving. *IEEE Trans. Robot.* **2018**, *34*, 1603–1622. [CrossRef]
12. Kschischang, F.R.; Frey, B.J.; Loeliger, H.A. Factor graphs and the sum-product algorithm. *IEEE Trans. Inf. Theory* **2001**, *47*, 498–519. [CrossRef]
13. Şenöz, İ.; van de Laar, T.; Bagaev, D.; de Vries, B. Variational message passing and local constraint manipulation in factor graphs. *Entropy* **2021**, *23*, 807. [CrossRef]
14. Hoffmann, C.; Rostalski, P. Linear optimal control on factor graphs—a message passing perspective. *IFAC-PapersOnLine* **2017**, *50*, 6314–6319. [CrossRef]
15. Loeliger, H.A.; Dauwels, J.; Hu, J.; Korl, S.; Ping, L.; Kschischang, F.R. The factor graph approach to model-based signal processing. *Proc. IEEE* **2007**, *95*, 1295–1322. [CrossRef]
16. Cox, M.; van de Laar, T.; de Vries, B. A factor graph approach to automated design of Bayesian signal processing algorithms. *Int. J. Approx. Reason.* **2019**, *104*, 185–204. [CrossRef]
17. Palmieri, F.A.; Pattipati, K.R.; Di Gennaro, G.; Fioretti, G.; Verolla, F.; Buonanno, A. A unifying view of estimation and control using belief propagation with application to path planning. *IEEE Access* **2022**, *10*, 15193–15216. [CrossRef]
18. Forney, G.D. Codes on graphs: Normal realizations. *IEEE Trans. Inf. Theory* **2001**, *47*, 520–548. [CrossRef]
19. Le, F.; Srivatsa, M.; Reddy, K.K.; Roy, K. Using graphical models as explanations in deep neural networks. In Proceedings of the IEEE International Conference on Mobile Ad-Hoc and Smart Systems, Monterey, CA, USA, 4–7 November 2019; pp. 283–289.
20. Lecue, F. On the role of knowledge graphs in explainable AI. *Semant. Web* **2020**, *11*, 41–51. [CrossRef]
21. Yedidia, J.S.; Freeman, W.T.; Weiss, Y. Bethe free energy, Kikuchi approximations, and belief propagation algorithms. *Adv. Neural Inf. Process. Syst.* **2001**, *13*. Available online: https://merl.com/publications/docs/TR2001-16.pdf (accessed on 8 June 2025).
22. Zhang, Y.; Xu, W.; Liu, A.; Lau, V. Message Passing Based Wireless Federated Learning via Analog Message Aggregation. In Proceedings of the IEEE/CIC International Conference on Communications in China, Hangzhou, China, 7–9 August 2024; pp. 2161–2166.
23. Bagaev, D.; de Vries, B. Reactive message passing for scalable Bayesian inference. *Sci. Program.* **2023**, *2023*, 6601690. [CrossRef]
24. Podusenko, A.; Kouw, W.M.; de Vries, B. Message passing-based inference for time-varying autoregressive models. *Entropy* **2021**, *23*, 683. [CrossRef]
25. Kouw, W.M.; Podusenko, A.; Koudahl, M.T.; Schoukens, M. Variational message passing for online polynomial NARMAX identification. In Proceedings of the American Control Conference, Atlanta, GA, USA, 8–10 June 2022; IEEE: New York, NY, USA, 2022; pp. 2755–2760.
26. Petersen, K.B.; Pedersen, M.S. The matrix cookbook. *Tech. Univ. Den.* **2008**, *7*, 510.
27. Soch, J.; Allefeld, C.; Faulkenberry, T.J.; Pavlovic, M.; Petrykowski, K.; Sarıtaş, K.; Balkus, S.; Kipnis, A.; Atze, H.; Martin, O.A. The Book of Statistical Proofs (Version 2023). 2024. Available online: https://zenodo.org/records/10495684 (accessed on 8 June 2025).
28. Gupta, A.K.; Nagar, D.K. *Matrix Variate Distributions*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2018.
29. Särkkä, S. *Bayesian Filtering and Smoothing*; Cambridge University Press: London, UK; New York, NY, USA, 2013.
30. Lopes, M.T.; Castello, D.A.; Matt, C.F.T. A Bayesian inference approach to estimate elastic and damping parameters of a structure subjected to vibration tests. In Proceedings of the Inverse Problems, Design and Optimization Symposium, Joao Pessoa, Brazil, 25–27 August 2010.
31. Winn, J.; Bishop, C.M.; Jaakkola, T. Variational message passing. *J. Mach. Learn. Res.* **2005**, *6*, 661–694.
32. Dauwels, J.; Korl, S.; Loeliger, H.A. Particle methods as message passing. In Proceedings of the IEEE International Symposium on Information Theory, Seattle, DC, USA, 9–14 July 2006; pp. 2052–2056.
33. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; MIT Press: Cambridge, MA, USA, 2012.
34. Smith, R.; Friston, K.J.; Whyte, C.J. A step-by-step tutorial on active inference and its application to empirical data. *J. Math. Psychol.* **2022**, *107*, 102632. [CrossRef] [PubMed]

35. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational inference: A review for statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859–877. [CrossRef]
36. Parr, T.; Pezzulo, G.; Friston, K.J. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*; MIT Press: Cambridge, MA, USA, 2022.
37. Friston, K.; Da Costa, L.; Sajid, N.; Heins, C.; Ueltzhöffer, K.; Pavliotis, G.A.; Parr, T. The free energy principle made simpler but not too simple. *Phys. Rep.* **2023**, *1024*, 1–29. [CrossRef]
38. Yedidia, J.S.; Freeman, W.T.; Weiss, Y. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. Inf. Theory* **2005**, *51*, 2282–2312. [CrossRef]
39. Proakis, J.G. *Digital Signal Processing: Principles Algorithms and Applications*; Pearson Education India: Noida, India, 2001.
40. Robertson, D.G.E.; Dowling, J.J. Design and responses of Butterworth and critically damped digital filters. *J. Electromyogr. Kinesiol.* **2003**, *13*, 569–573. [CrossRef]
41. Smith, J.O. *Introduction to Digital Filters: With Audio Applications*; Smith, J., Ed.; W3K Publishing: San Francisco, CA, USA, 2008; Volume 2.
42. Zumbahlen, H. (Ed.) *Linear Circuit Design Handbook*; Newnes: Oxford, UK, 2011.
43. Mello, R.G.; Oliveira, L.F.; Nadal, J. Digital Butterworth filter for subtracting noise from low magnitude surface electromyogram. *Comput. Methods Programs Biomed.* **2007**, *87*, 28–35. [CrossRef]
44. Damgaard, M.R.; Pedersen, R.; Bak, T. Study of variational inference for flexible distributed probabilistic robotics. *Robotics* **2022**, *11*, 38. [CrossRef]
45. Tedeschini, B.C.; Brambilla, M.; Nicoli, M. Message passing neural network versus message passing algorithm for cooperative positioning. *IEEE Trans. Cogn. Commun. Netw.* **2023**, *9*, 1666–1676. [CrossRef]
46. Ta, D.N.; Kobilarov, M.; Dellaert, F. A factor graph approach to estimation and model predictive control on unmanned aerial vehicles. In Proceedings of the International Conference on Unmanned Aircraft Systems, Orlando, FL, USA, 27–30 May 2014; IEEE: New York, NY, USA, 2014; pp. 181–188.
47. Castaldo, F.; Palmieri, F.A. A multi-camera multi-target tracker based on factor graphs. In Proceedings of the IEEE International Symposium on Innovations in Intelligent Systems and Applications, Alberobello, Italy, 23–25 June 2014; IEEE: New York, NY, USA, 2014; pp. 131–137.
48. van Erp, B.; Bagaev, D.; Podusenko, A.; Şenöz, İ.; de Vries, B. Multi-agent trajectory planning with NUV priors. In Proceedings of the American Control Conference, Toronto, ON, Canada, 10–12 July 2024; IEEE: New York, NY, USA, 2024; pp. 2766–2771.
49. Assimakis, N.; Adam, M.; Douladiris, A. Information filter and Kalman filter comparison: Selection of the faster filter. In Proceedings of the Information Engineering, Chongqing, China, 26–28 October 2012; Volume 2, pp. 1–5.
50. Cover, T.M. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 1999.