Variational message passing for online polynomial NARMAX identification

Wouter Kouw, Albert Podusenko, Magnus Koudahl & Maarten Schoukens

American Control Conference 2022

Dept. EE, Bayesian Intelligent Autonomous Systems lab

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

NARMAX Model

Nonlinear autoregressive moving average with exogenous input.

- Important model class for system identification, e.g., robot arm control (Billings, 2013).
- Polynomial NARMAX:

$$y_k = \theta^\top \phi(u_k, u_{k\text{-}1} \ldots u_{k\text{-}n_a}, y_{k\text{-}1} \ldots y_{k\text{-}n_b}, e_{k\text{-}1} \ldots e_{k\text{-}n_e}) + e_k$$

where θ are coefficients and ϕ is a polynomial basis expansion.

The noise instances e_k follow a zero-mean Gaussian distribution:

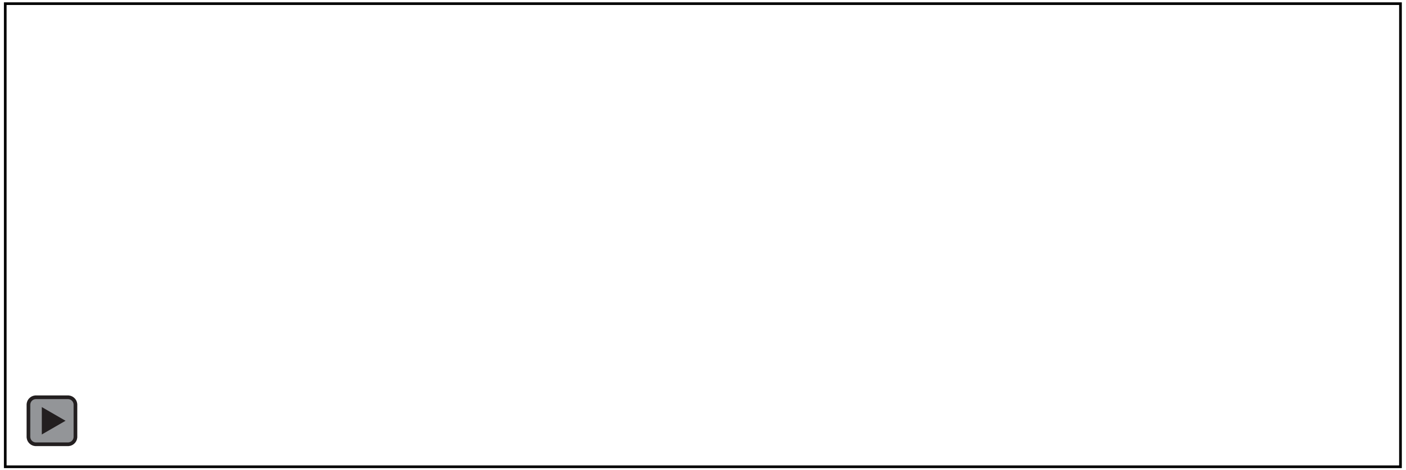
$$e_k \sim \mathcal{N}(0, \tau^{-1})$$

where τ is a precision (inverse variance) parameter.

Bayesian Inference

What is Bayesian inference?

- Updating a belief over an unknown variable in a model given data.



Why Bayesian Inference?

Point estimators (e.g., least-squares, max-likelihood) overfit when sample size is low.

- Overfitting: fitting to noise in training data, which deteriorates predictions.

Bayesian inference is naturally robust to overfitting.

- The use of probability distributions has a naturally regularizing effect, avoiding noise samples.

Why isn’t Bayesian inference used everywhere?

- It is typically computationally expensive to run the inference algorithm.

Solution: variational Bayesian inference distributed over a factor graph.

Probabilistic NARMAX Model

By absorbing the noise e_k, we can cast the dynamics as a Gaussian likelihood:

$$p(y_k \mid u_k, \mathbf{u}_{k\text{-}1}, \mathbf{y}_{k\text{-}1}, \mathbf{e}_{k\text{-}1}, \theta, \tau) = \mathcal{N}\big(y_k \mid \theta^\top \phi_k, \tau^{-1}\big)$$

where the bold symbols are the vectors of delayed inputs, outputs and noises,

$$\mathbf{u}_{k\text{-}1} = [u_{k\text{-}1} \,..\, u_{k\text{-}n_a}] \quad \mathbf{y}_{k\text{-}1} = [y_{k\text{-}1} \,..\, y_{k\text{-}n_b}] \quad \mathbf{e}_{k\text{-}1} = [e_{k\text{-}1} \,..\, e_{k\text{-}n_e}]$$

and ϕ_k is shorthand for

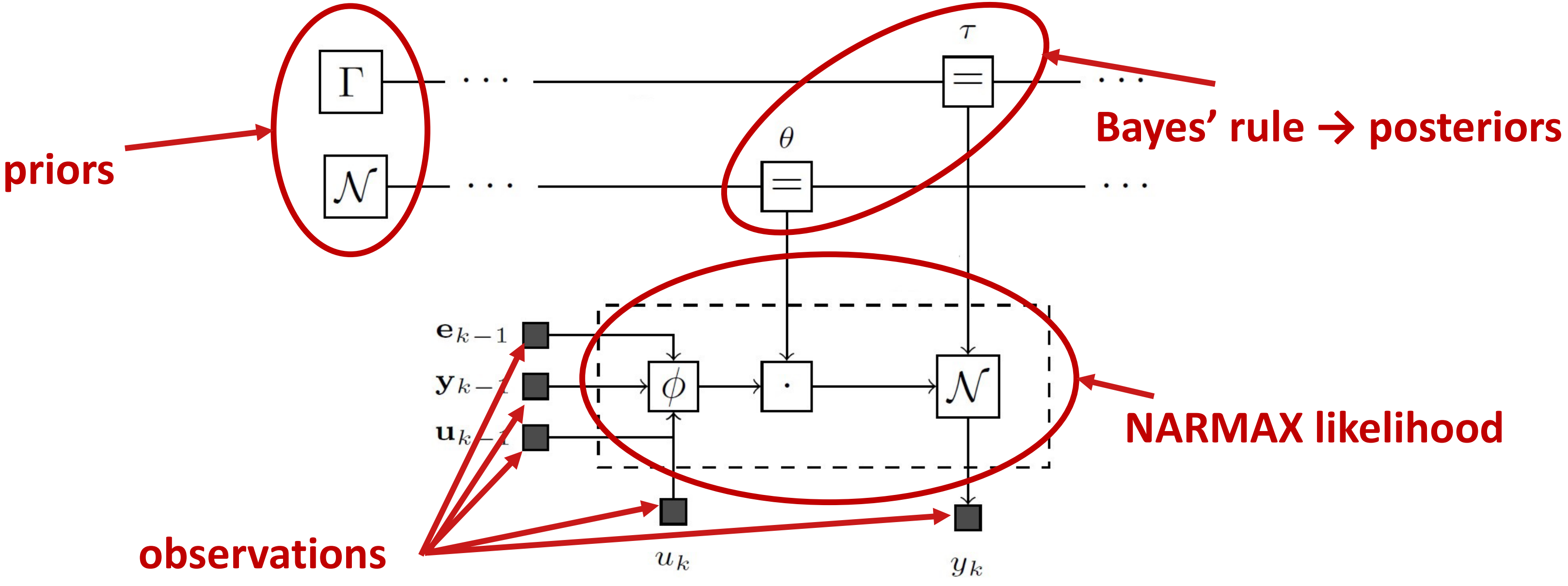
$$\phi_k = \phi(u_k, \mathbf{u}_{k\text{-}1}, \mathbf{y}_{k\text{-}1}, \mathbf{e}_{k-1})$$

The coefficients θ and noise precision τ are unknown and require prior distributions.

- Prior distribution for coefficients: $p(\theta) = \mathcal{N}(\theta \mid \mu_0, \Lambda_0^{-1})$
- Prior distribution for noise: $p(\tau) = \mathcal{G}(\tau \mid \alpha_0, \beta_0)$

Factor graph

We can map factors of the probabilistic model to nodes in a graph:



Variational message passing for online Bayesian NARMAX identification (Kouw, Podusenko, Koudahl & Schoukens).

Bayesian filtering

We derive a recursive expression for the posterior distributions:

- At $k = 1$, we have:

posterior evidence likelihood priors

$$p(\theta, \tau \mid y_1, u_1) = \frac{1}{p(y_1 \mid u_1)} p(y_1 \mid u_1, \theta, \tau) p(\theta) p(\tau)$$

- At $\mathrm{k} = 2$, we have the first appearance of a previous noise instance:

$$p(\theta, \tau \mid y_{1:2}, u_{1:2}, e_1) = \frac{1}{p(y_2 \mid u_{1:2}, y_1, e_1)} p(y_2 \mid u_{1:2}, y_1, e_1, \theta, \tau) p(\theta, \tau \mid y_1, u_1)$$

- At $\mathrm{k} > 2$, we update the posterior recursively based on incoming data:

$$p(\theta, \tau \mid y_{1:k}, u_{1:k}, e_{1:k\text{-}1}) = \frac{p(y_k \mid u_k, \mathbf{u}_{k\text{-}1}, \mathbf{y}_{k\text{-}1}, \mathbf{e}_{k\text{-}1}, \theta, \tau)}{p(y_k \mid u_{1:k}, y_{1:k\text{-}1}, e_{1:k\text{-}1})} p(\theta, \tau \mid y_{1:k\text{-}1}, u_{1:k\text{-}1}, e_{1:k\text{-}2})$$

Inference

We use variational Bayes to infer coefficients and noise simultaneously.

- Form a free energy objective function with respect to a recognition model q:

$$\mathcal{F}_k[q_k] = \iint q_k(\theta,\tau) \ln \frac{q_k(\theta,\tau)}{p(y_k,\theta,\tau \mid u_{1:k}, y_{1:k\text{-}1}, e_{1:k\text{-}1})} \mathrm{d}\theta \mathrm{d}\tau$$

This free energy function can be decomposed into more easily computable terms.

The recognition model is chosen to be:

- For coefficients: $q_k(\theta) = \mathcal{N}(\theta \mid \mu_k, \Lambda_k^{-1})$
- For precision: $q_k(\tau) = \mathcal{G}(\tau \mid \alpha_k, \beta_k)$

Update equations

Optimal forms of the marginal recognition factors can be derived:

- For coefficients:

$$q_k(\theta) \propto \exp\Big(\mathbb{E}_{q_k(\tau)} \ln p(\theta, \tau \mid y_{1:k\text{-}1}, u_{1:k\text{-}1}, e_{1:k\text{-}2}) \\ + \mathbb{E}_{q_k(\tau)} \ln p(y_k \mid u_k, \mathbf{u}_{k\text{-}1}, \mathbf{y}_{k\text{-}1}, \mathbf{e}_{k\text{-}1}, \theta, \tau)\Big)$$

- For precision:

$$q_k(\tau) \propto \exp\Big(\mathbb{E}_{q_k(\theta)} \ln p(\theta, \tau \mid y_{1:k\text{-}1}, u_{1:k\text{-}1}, e_{1:k\text{-}2}) \\ + \mathbb{E}_{q_k(\theta)} \ln p(y_k \mid u_k, \mathbf{u}_{k\text{-}1}, \mathbf{y}_{k\text{-}1}, \mathbf{e}_{k\text{-}1}, \theta, \tau)\Big)$$

We can map terms of the marginal updates to (variational) messages on the factor graph.

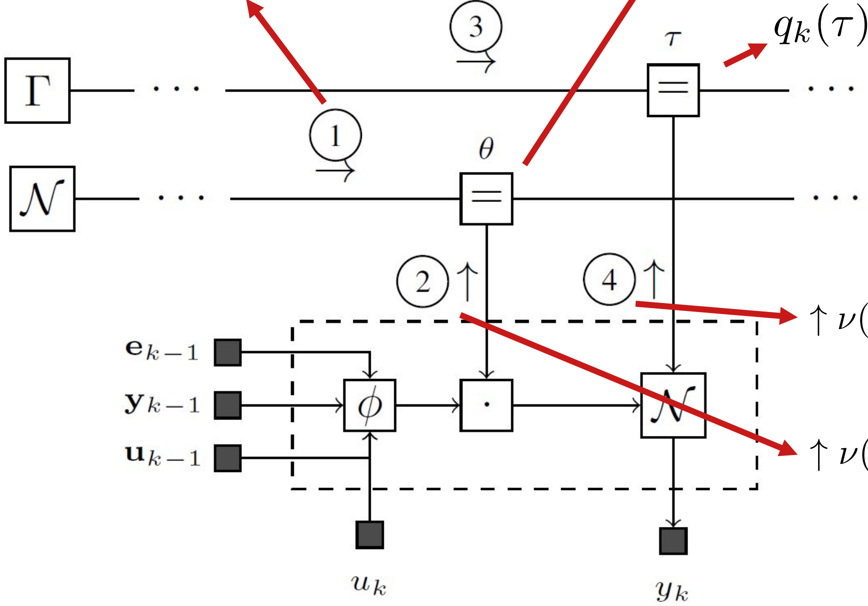
Variational message passing

$$\overrightarrow{\nu}(\tau) \propto \mathcal{G}(\tau \mid \alpha_{k\text{-}1}, \beta_{k\text{-}1})$$

$$\overrightarrow{\nu}(\theta) \propto \mathcal{N}(\theta \mid \mu_{k\text{-}1}, \Lambda_{k\text{-}1}^{-1})$$

$$q_k(\theta) = \mathcal{N}(\theta \mid \Lambda_k \mu_k = \Lambda_{k\text{-}1}\mu_{k\text{-}1} + \bar{\Lambda}_k \bar{\mu}_k, \Lambda_k = \Lambda_{k\text{-}1} + \bar{\Lambda}_k)$$

$$q_k(\tau) = \mathcal{G}(\tau \mid \alpha_k = \alpha_{k\text{-}1} + \bar{\alpha}_k,\ \beta_k = \beta_{k\text{-}1} + \bar{\beta}_k)$$



$$\uparrow \nu(\tau) \propto \mathcal{G}(\tau \mid \bar{\alpha}_k = \frac{3}{2}, \bar{\beta}_k = \frac{1}{2}((y_k - \mu_k^\top \phi_k)^2 + \phi_k^\top \Lambda_k^{-1} \phi_k))$$

$$\uparrow \nu(\theta) \propto \mathcal{N}(\theta \mid \bar{\Lambda}_k \bar{\mu}_k = \frac{\alpha_k}{\beta_k} y_k \phi_k, \bar{\Lambda}_k = \frac{\alpha_k}{\beta_k} \phi_k \phi_k^\top)$$

 Variational message passing for online Bayesian NARMAX identification (Kouw, Podusenko, Koudahl & Schoukens).

Posterior predictive

Goal of system identification is to generate outputs for given inputs using the model.

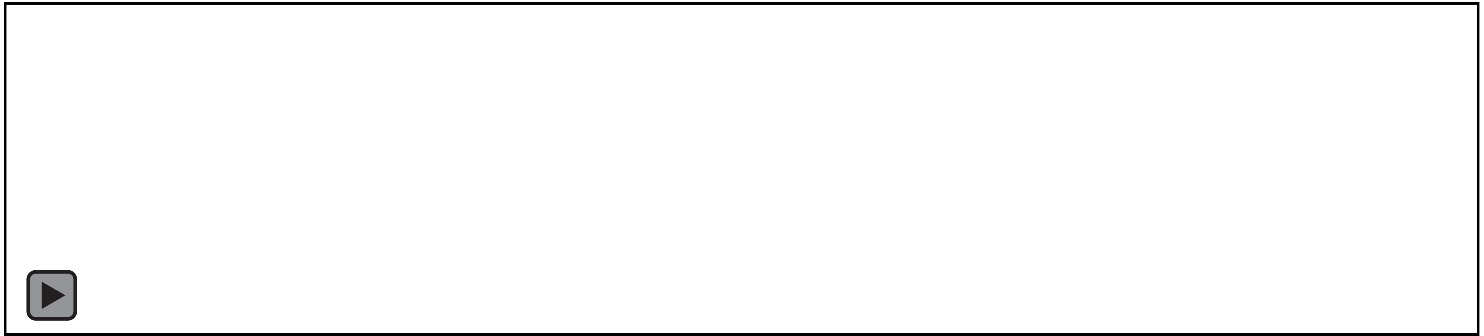
- Unobserved outputs are random variables → Bayesian inference.

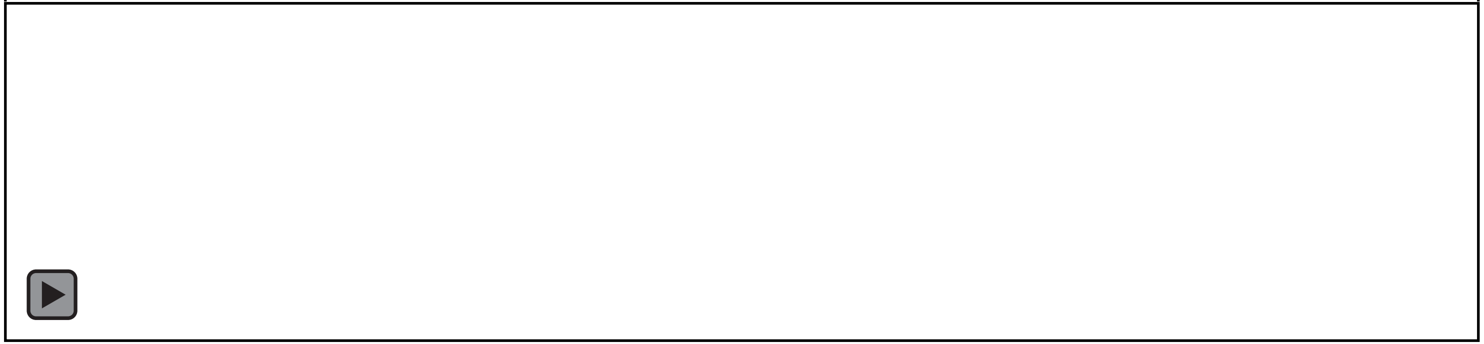
Posterior predictive distribution for future output:

$$p(y_j \mid u_j, u_{1:T}, y_{1:T}, e_{1:T}) \approx \mathcal{N}\big(y_j \mid \mu_T^\top \phi_j,\ \phi_j^\top \Lambda_T^{-1} \phi_j + \tfrac{\beta_T}{\alpha_T}\big)$$

where $\mu_T, \Lambda_T, \alpha_T, \beta_T$ refer to the parameters of the recognition factors at $k = T$.

Demo: simulation

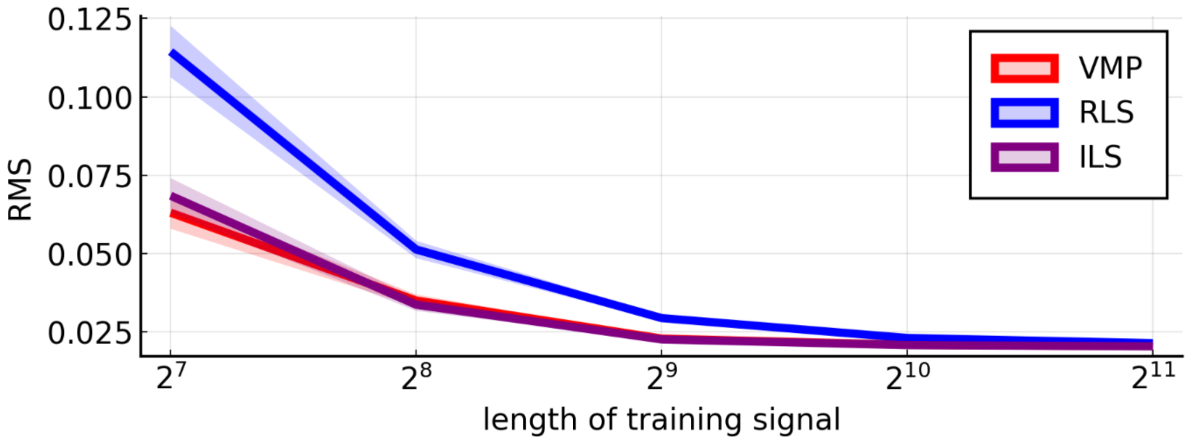




Experimental results

100 Monte Carlo runs of a verification experiment:

- Polynomial order = 3, delays = 1, coefficients pseudo-randomly sampled.



Conclusion

Variational Bayes outperforms least-squares in small sample size regimes.

- Prior distributions need not be informative.

Variational message passing is an efficient inference algorithm.

- At 2 iterations, the computational complexity is 4 times that of RLS.

Questions?

Collaborators:

Albert Podusenko

Magnus Koudahl

Maarten Schoukens

Extra slide: previous noise instances

At k, we had previous noise instances $\boldsymbol{e}_{k-1}$ that were treated as observed variables.

We populate this vector as follows:

1. At $k-1$, we compute a posterior predictive distribution for e_k.
2. That distribution is collapsed to a point estimate (specifically, the MAP).
3. We transition to k, observe y_k and calculate prediction error:

$$e_k \triangleq y_k - \mu_{k-1}^{\top} \phi_k$$

4. When transitioning to $k+1$, this prediction error is added to $\boldsymbol{e}_{k-1}$ and the oldest entry is dropped.

Note that this entails information loss -> we have a new paper that avoids this.

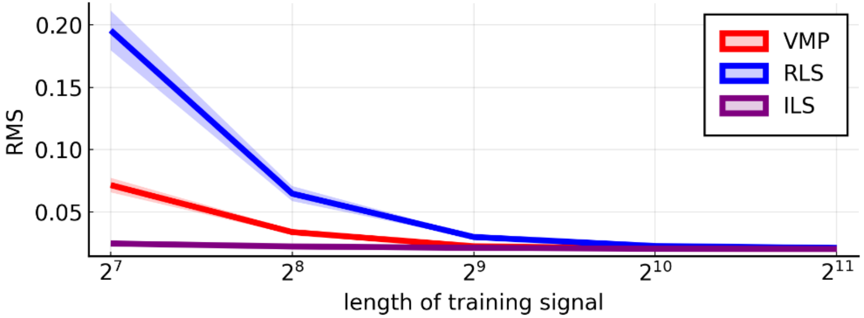
 Variational message passing for online Bayesian NARMAX identification (Kouw, Podusenko, Koudahl & Schoukens).

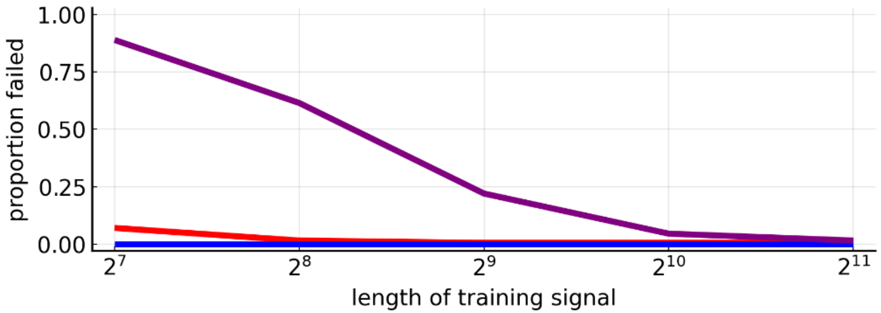
Extra slide: consistency of coefficient posterior

Coefficient posterior concentrates on true parameters:

Extra slide: 1-step ahead prediction

Verification experiment with 1-step ahead predictions:

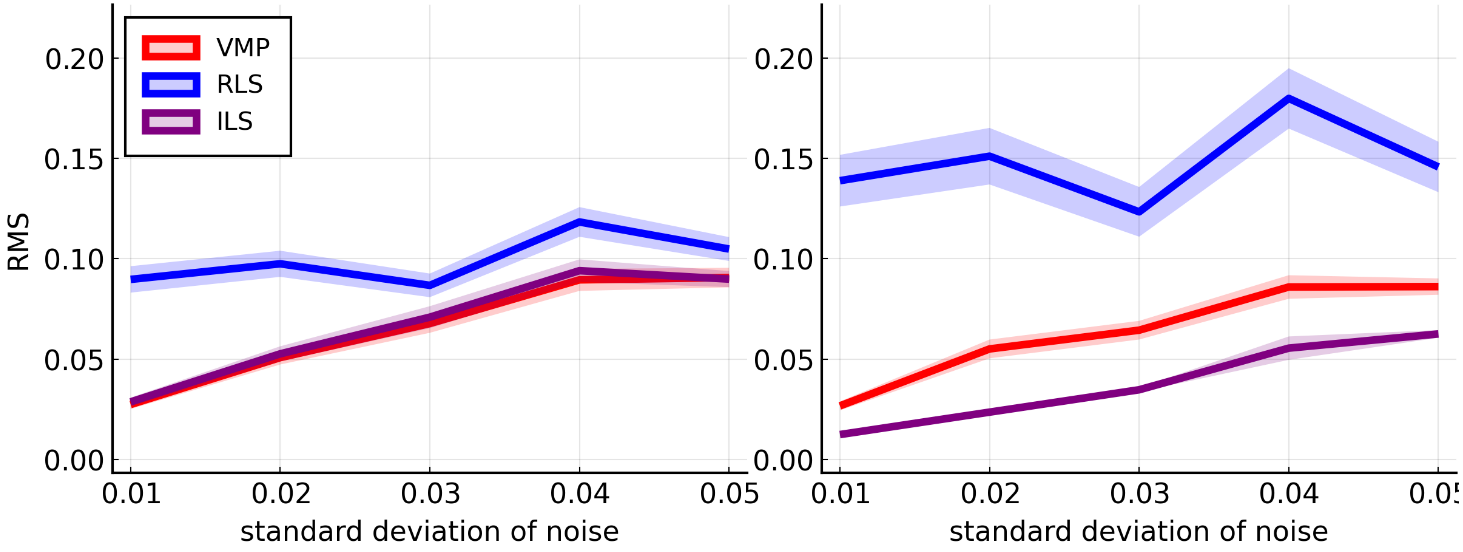




ILS fails much more often due to unstable parameter estimates.

Extra slide: system noise sweep experiment

Verification experiment with varying system noise parameter:



TU/e