

Variational Bayesian inference for system identification

Wouter M. Kouw
Nonlinear System Identification Workshop 2023

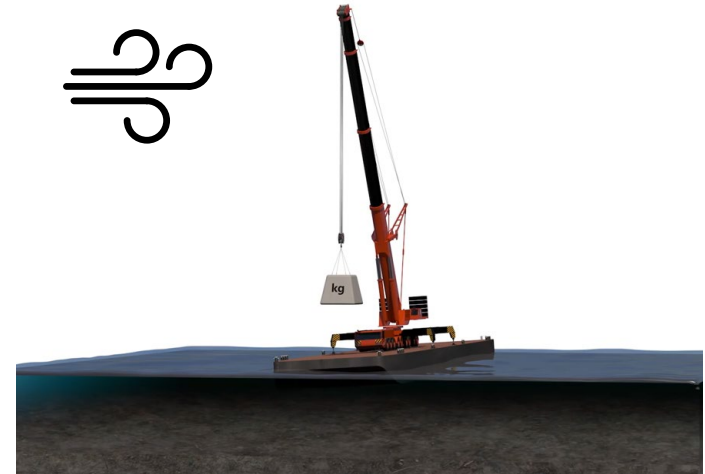
<https://github.com/wmkouw/NSIW2023-keynote>



Source:
Dick van Smirren

Uncertainty

- What value do parameters in my model have?
- How many parameters affect my system?
- Do my parameters change over time?
- Is my system affected by external disturbances?
- Which model should I select for this system?
- What should I measure to identify my system?



Modelling

Typical models for system identification look something like

$$\begin{aligned}x_k &= f_{\theta}(x_{k-1}, u_k) + w_k , \\ y_k &= g_{\eta}(x_k) + v_k .\end{aligned}$$

or like

$$y_k = f_{\theta}(u_k, u_{k-1}, \dots, y_{k-1}, \dots) + e_k ,$$

But where are the uncertainties?

Probabilistic modelling

Probabilistic models aim to include more sources of uncertainty:

$$p(y, u, \theta, \sigma) = p(y|u, \theta, \sigma) p(u) p(\theta) p(\sigma)$$

generative model

observation model

input prior

parameter priors

Formally, one also conditions on assumptions leading to model design:

$$p(y, u, \theta, \sigma \mid \underline{\mathcal{M} = m_1})$$

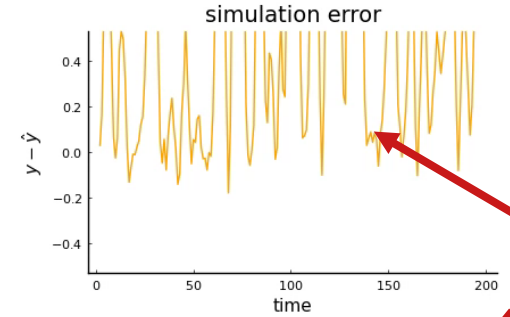
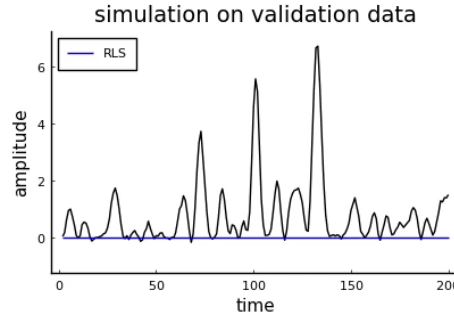
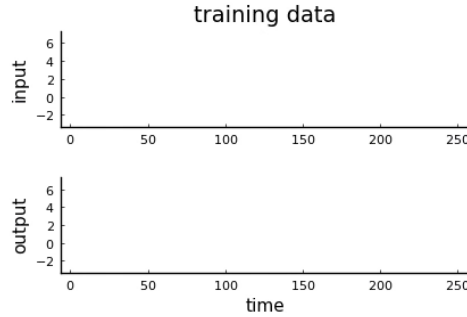
Inference

We can estimate unknowns by inverting the model:

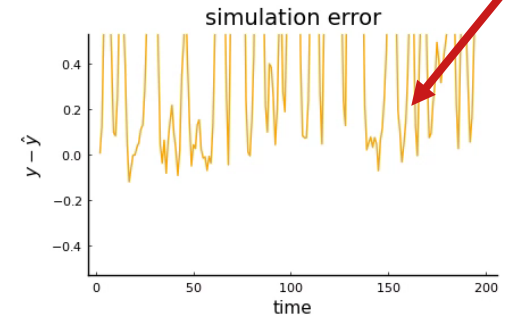
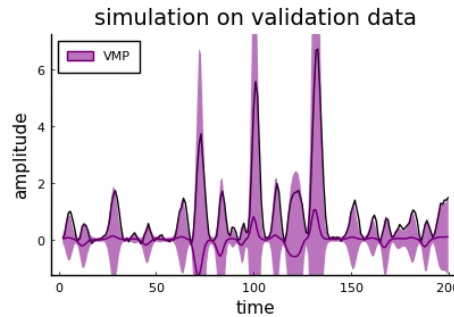
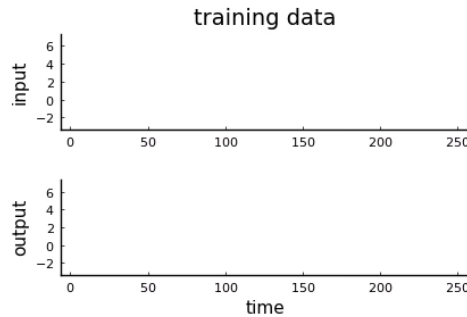
$$\underbrace{p(\theta | x)}_{\text{posterior distribution}} = \frac{\overbrace{p(x | \theta)}^{\text{likelihood}}}{\underbrace{p(x)}_{\text{evidence}}} \underbrace{p(\theta)}_{\text{prior distribution}}$$

This is known as Bayes' rule.

Is all that extra work useful?

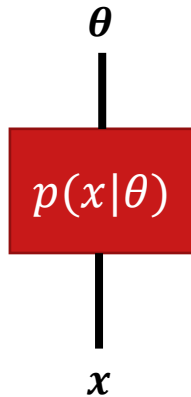


Yes



Factor graphs

Probabilistic model equations quickly become complex and hard to read.
It helps to adopt a visual language: factor graphs.

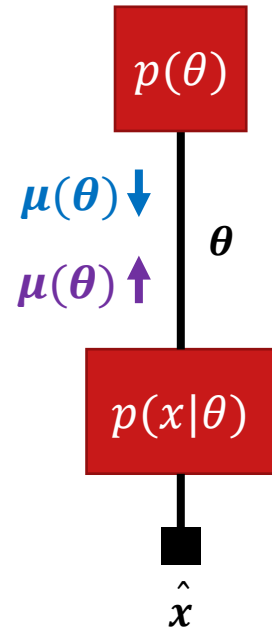


Edges represent variables in the model.

Nodes represent relationships between variables.

Message passing

The following is a complete factor graph:



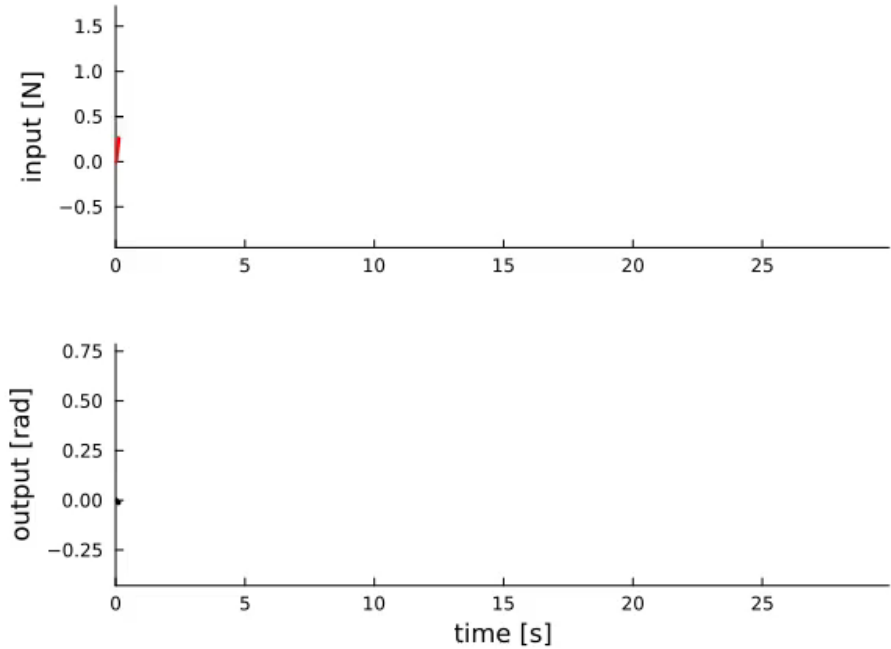
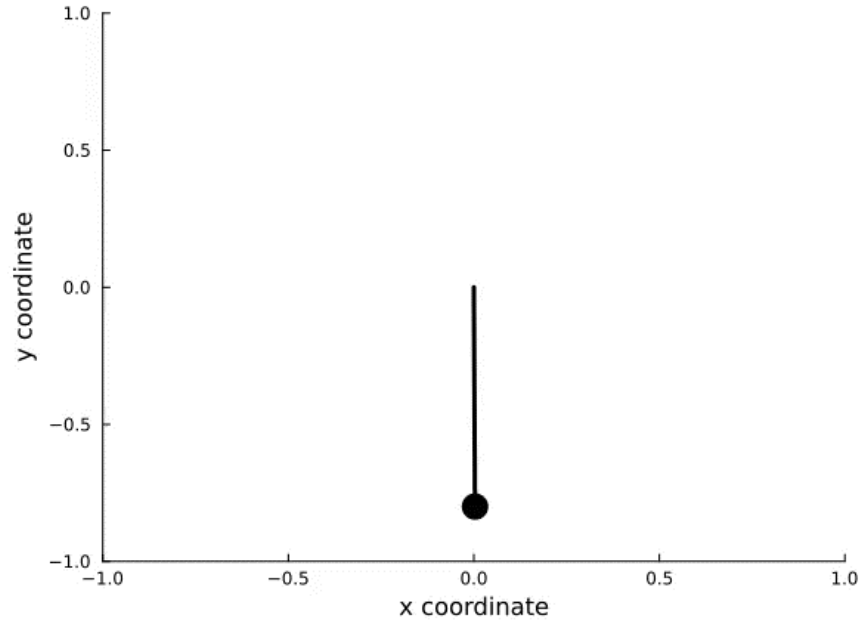
Terminal nodes are priors.

The combination of the prior and the likelihood to form the posterior can be expressed as messages passed from nodes.

$$p(\theta|x = \hat{x}) \propto \int \delta(x - \hat{x}) p(x|\theta) dx p(\theta)$$

Black nodes represent observed data.

Demonstration system



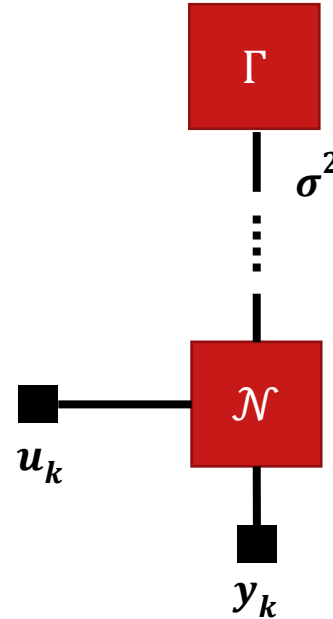
Model 1

Consider a prediction based on an unaltered input u_k with likelihood variance σ^2 :

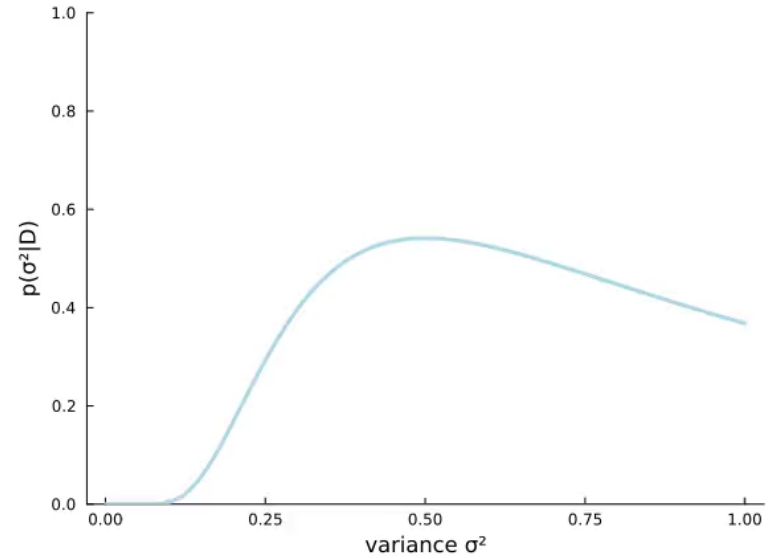
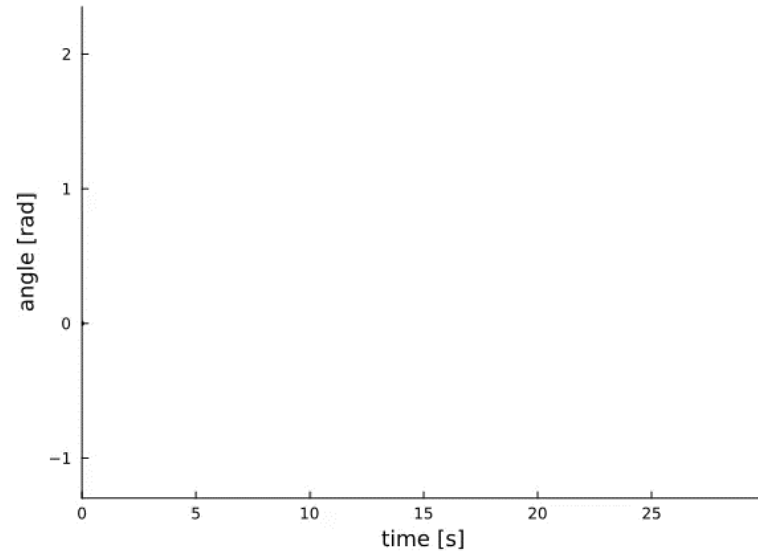
$$y_k = u_k + e_k, \quad \text{with } e_k \sim \mathcal{N}(0, \sigma^2) .$$

In probabilistic model form, this could become:

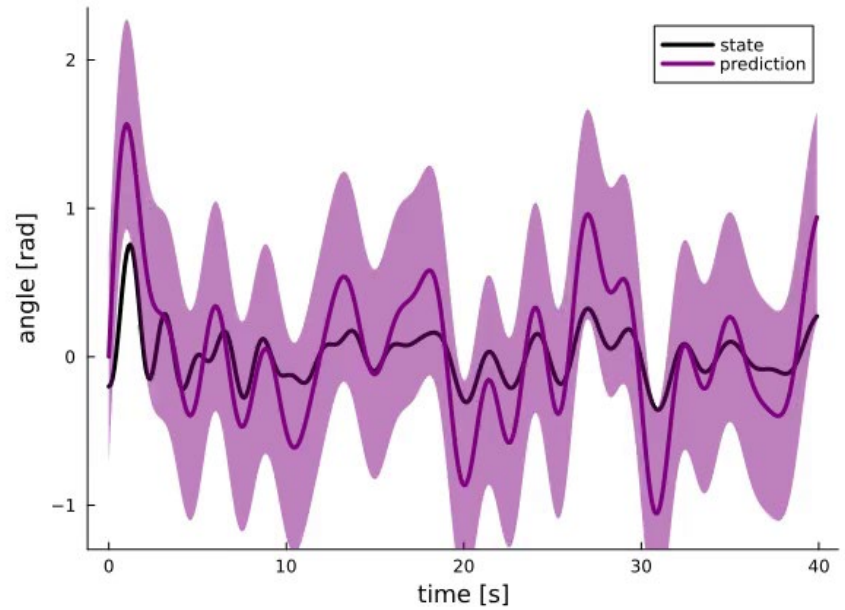
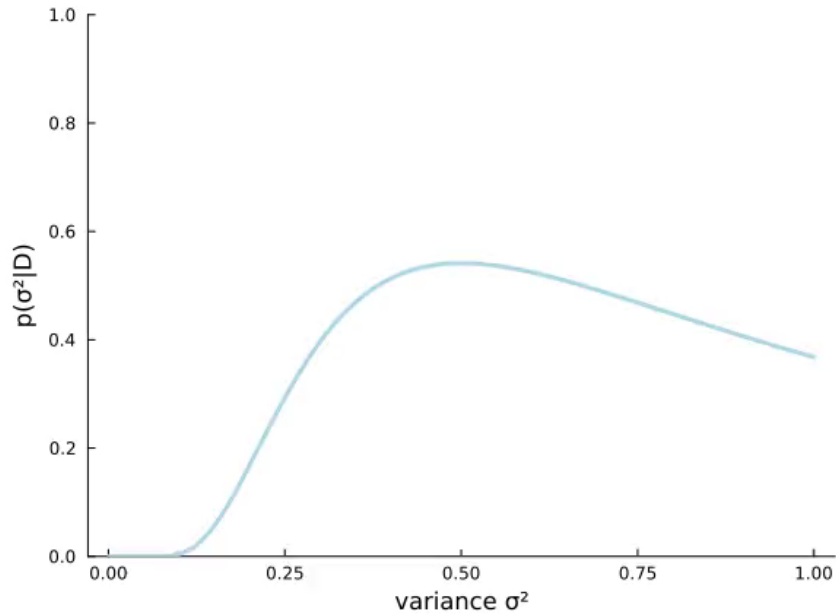
$$p(y_k, \sigma^2 | u_k) = \mathcal{N}(y_k | u_k, \sigma^2) \Gamma(\sigma^2 | \alpha, \beta) .$$



Model 1



Model 1



Model 1

This model obviously doesn't work very well.

A straightforward extension is a NARX model:

$$y_k = \theta^\top \varphi(u_k, u_{k-1}, \dots, y_{k-1}, \dots) + e_k ,$$

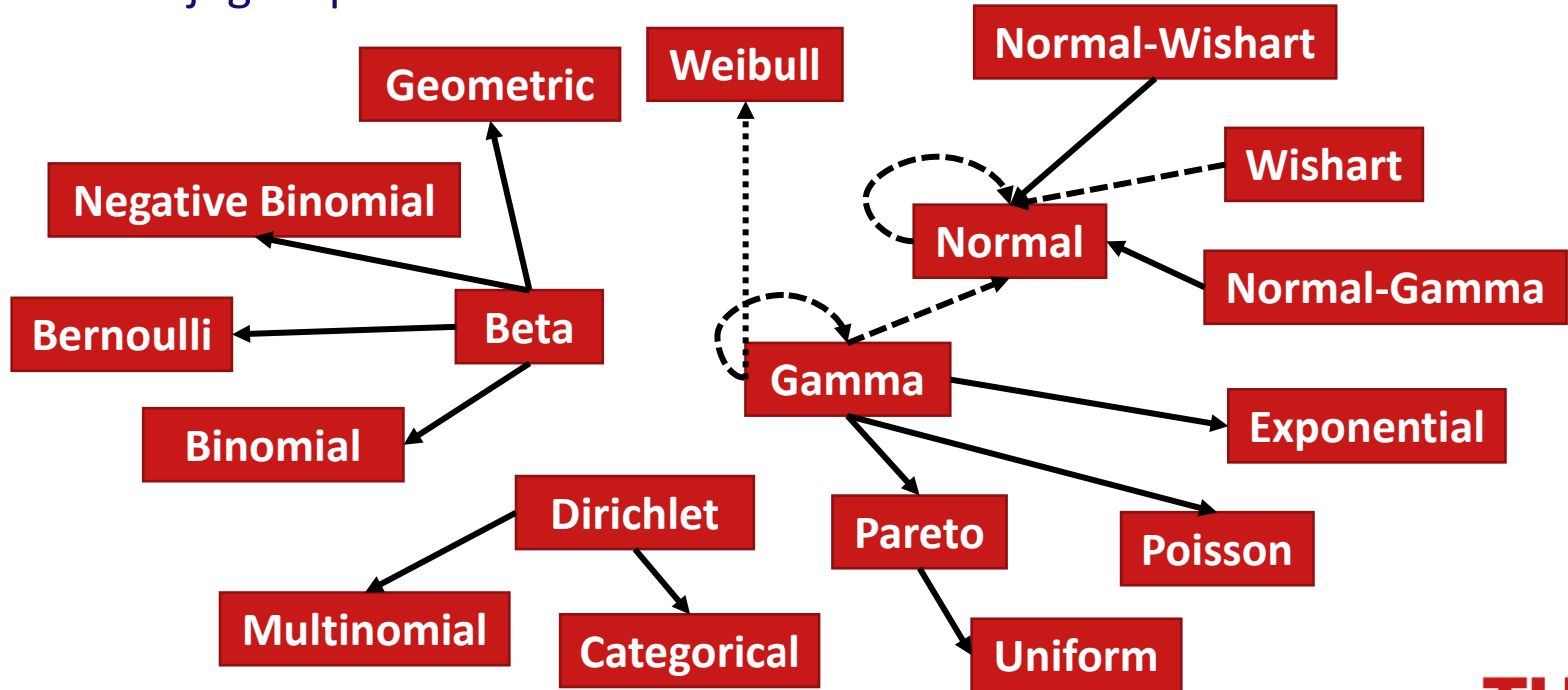
But now we run into a problem: we can't obtain a posterior distribution.

It requires solving an intractable integral:

$$p(y_k|u_k) = \iint p(y_k|u_k, \theta, \sigma^2) p(\theta) p(\sigma^2) d\theta d\sigma^2$$

Exact inference

Limited to conjugate priors:



Approximate inference

We may approximate the posterior $p(\theta|x)$ with a distribution $q(\theta)$.

To do that, we need an objective characterizing the dissimilarity between q and p .

$$\mathcal{F}[q] = \int_{\theta} q(\theta) \log \frac{q(\theta)}{p(\theta, x)} d\theta$$

This is known as a “free energy” functional and may be understood through:

$$\mathcal{F}[q] = \underbrace{\int_{\theta} q(\theta) \log \frac{1}{p(x|\theta)} d\theta}_{\text{prediction error}} + \underbrace{\int_{\theta} q(\theta) \log \frac{q(\theta)}{p(\theta)} d\theta}_{\text{complexity}}$$

Minimizing free energy

The free energy is a functional, i.e., a function of functions.

We are looking for the *probability distribution function* that minimizes it:

$$q^* = \arg \min_{q \in \mathcal{Q}} \mathcal{F}[q]$$

The space \mathcal{Q} represents the space of candidate functions.

Possible constraints on \mathcal{Q} include:

1. Data, $q(x) = \delta(x - \hat{x})$.
2. Parametrization, $q(\theta) = \mathcal{N}(\theta|m, v)$.
3. Factorization, $q(x, \theta) = q(x)q(\theta)$.
4. Probability mass in a subspace.

Minimizing free energy

Suppose we have a distribution $p(\theta)$ and we wish to minimize:

$$\mathcal{F}[q] = \int_{\Theta} q(\theta) \log \frac{q(\theta)}{p(\theta)} d\theta$$

The function q is constrained to be a valid probability distribution:

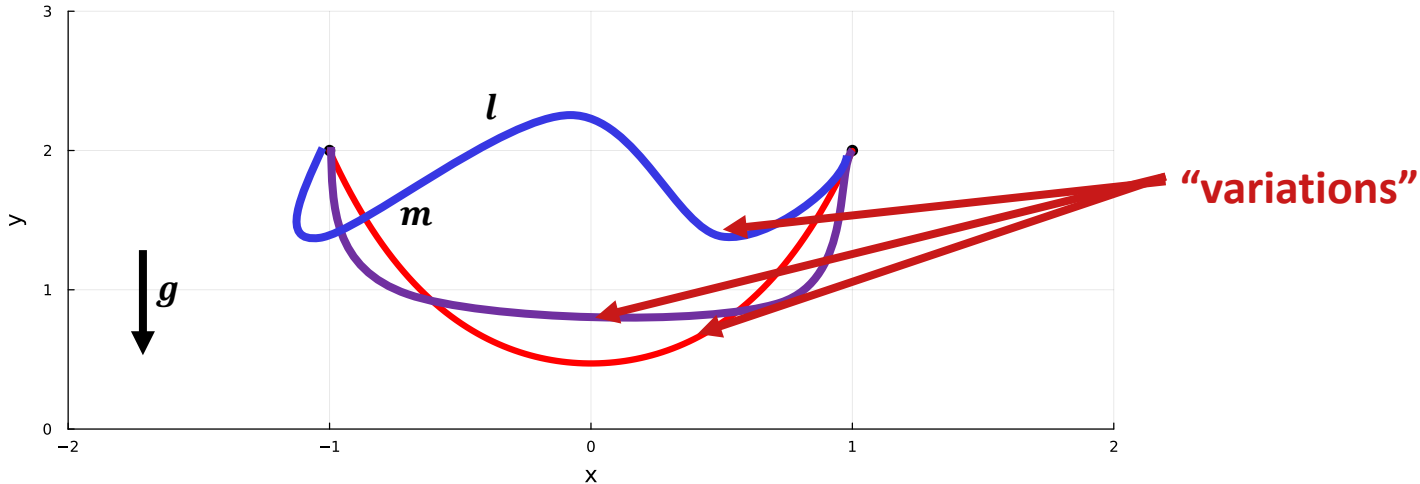
$$\mathcal{L}[q] = \mathcal{F}[q] + \lambda \left(\int_{\Theta} q(\theta) d\theta - 1 \right).$$

To find the minimizer, we must find the functional derivative $\frac{\delta}{\delta q} \mathcal{L}[q]$ and set it to 0.

In essence, variational Bayes turns integration into optimization.

Variations on a curve

Consider two fixed anchor points with a chain hanging between them:



The red chain minimizes *potential energy* (from Lagrangian mechanics).

In our probabilistic model, we have variations $q(\theta) = q^*(\theta) + \varepsilon\phi(\theta)$.

Minimizing free energy

We can find the functional derivative by considering how much the Lagrangian changes as a function of the variation, and setting that to 0;

$$\left. \frac{d}{d\varepsilon} \mathcal{L}[q^* + \varepsilon\phi] \right|_{\varepsilon=0} = 0$$

Expanding the Lagrangian gives:

$$\int_{\Theta} \frac{d}{d\varepsilon} (q^* + \varepsilon\phi) \log \frac{q^* + \varepsilon\phi}{p} \Big|_{\varepsilon=0} d\theta + \lambda \int_{\Theta} \frac{d}{d\varepsilon} (q^* + \varepsilon\phi) \Big|_{\varepsilon=0} d\theta = 0$$
$$\int_{\Theta} \left(\log \frac{q^*}{p} + 1 + \lambda \right) \phi d\theta = 0$$

Minimizing free energy

The common term is the functional derivative we were looking for.

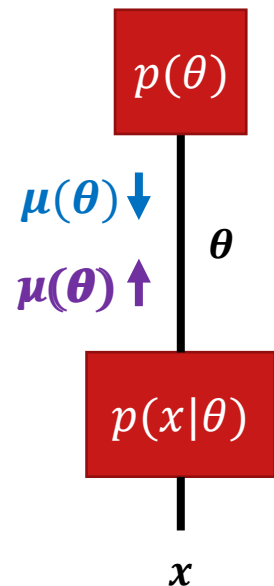
$$\int_{\Theta} \left(\log \frac{q^*}{p} + 1 + \lambda \right) \phi d\theta$$

The Lagrangian is 0 when the functional derivative is 0:

$$\frac{\delta}{\delta q} \mathcal{L}[q] = \log \frac{q^*}{p} + 1 + \lambda = 0$$
$$q^* = \frac{1}{\exp(1 + \lambda)} p$$

Variational message passing

One can distribute the free energy functional over a factor graph.

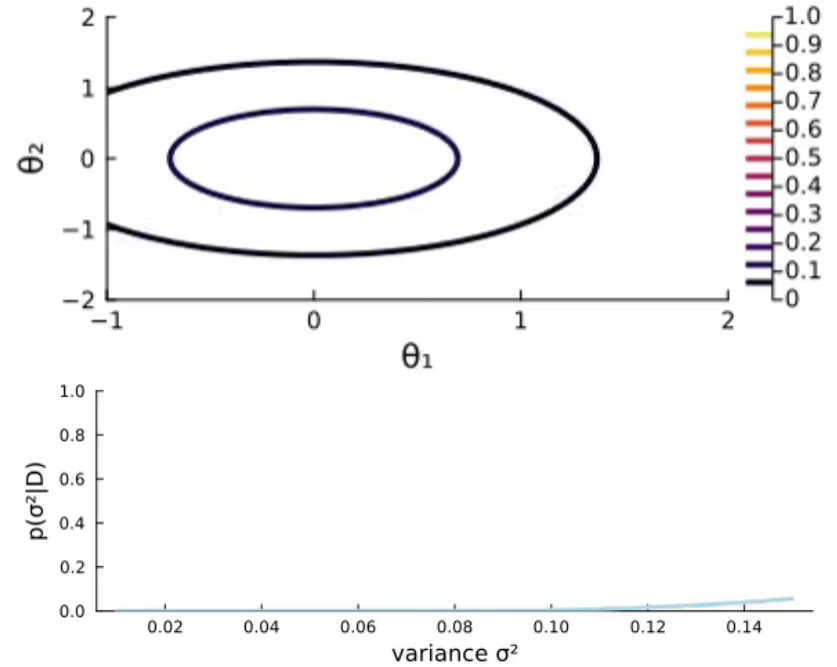
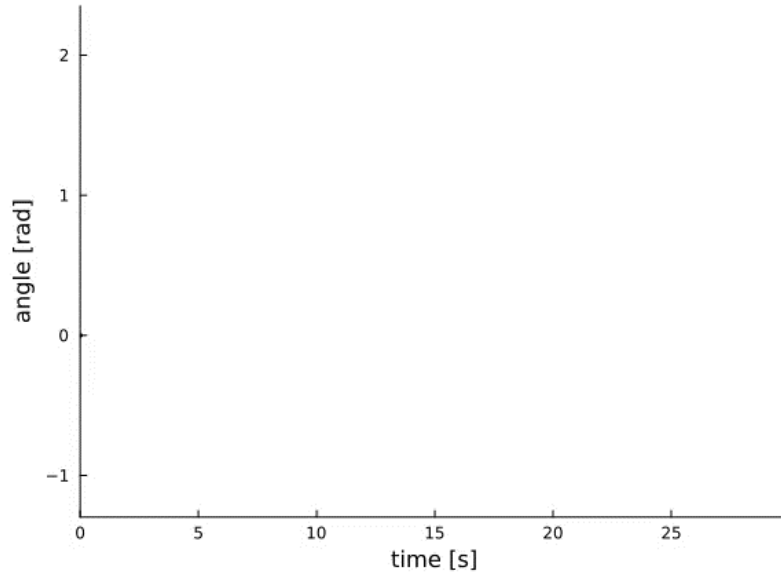


Variational approximation can be applied to factor nodes locally.

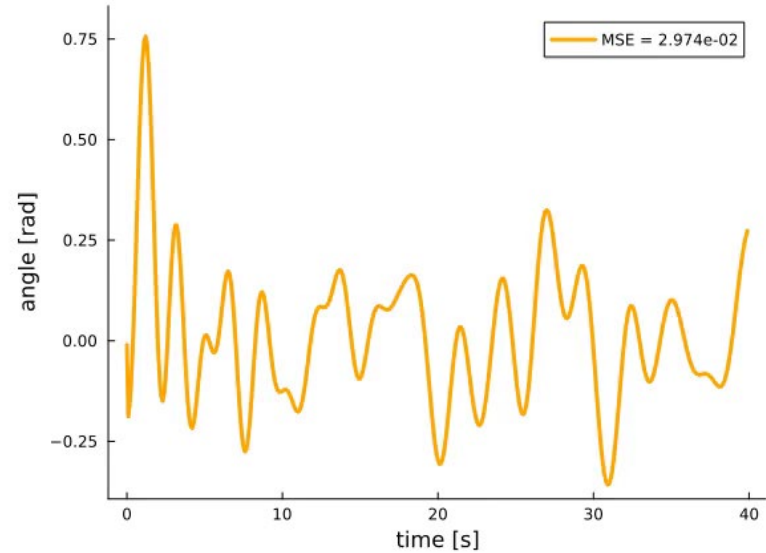
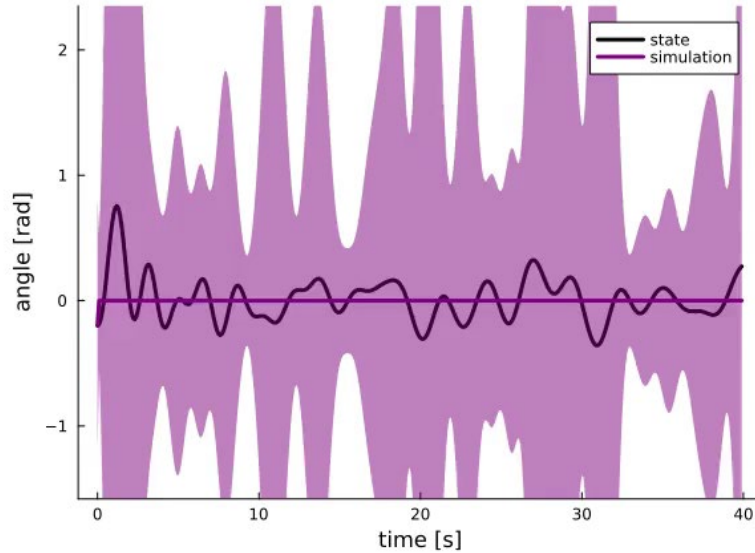
This turns standard messages into “variational messages”.

$$v(\theta) \propto \exp \left(\int_{\mathbf{x}} q(\mathbf{x}) \log p(\mathbf{x}|\theta) d\mathbf{x} \right)$$

Model 2



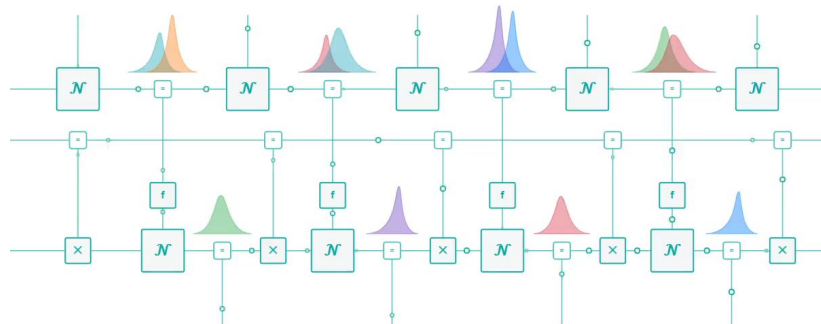
Model 2



Take-aways

1. Quantified uncertainty should be part of models.
2. Variational Bayes turns integration into optimization.
3. Variational message passing is inference distributed over a factor graph.

Checkout:  **rxinfer**



<https://github.com/biaslab/RxInfer.jl>



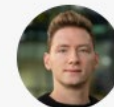
Bert de Vries
Professor, TU Eindhoven



Thijs van de Laar
Assistant professor, TU Eindhoven



Wouter Kouw
Assistant professor, TU Eindhoven



Albert Podusenko
Postdoctoral fellow, TU Eindhoven



Ismail Senoz
Postdoctoral fellow, TU Eindhoven



Magnus Tønder Koudahl
PhD candidate, TU Eindhoven



Chengfeng Jia
Guest PhD candidate, TU Eindhoven, WHUT



Dmitry V. Bagaev
PhD candidate, TU Eindhoven



Bart van Erp
PhD candidate - Teaching Assistant, TU Eindhoven



Tim Nisslbeck
PhD candidate, TU Eindhoven



Hoang Minh Huu Nguyen
PhD candidate, TU Eindhoven



Sepideh Adamiat
PhD candidate, TU Eindhoven



Mykola Lukashchuk
PhD candidate, TU Eindhoven



Wouter Nuijten
PhD candidate, TU Eindhoven



Martin Roa Villegas
PhD candidate, TU Eindhoven



Xianbo Xu
MSc Student, TU Eindhoven

Weakly informative priors

A common critique is that the act of “choosing priors” leads to non-objective results.

-> One should rely on as generic and uninformative priors as possible.

In the case of polynomial NARX models, I argue that one may use “weak information” in the sense that lower-order terms are more likely to have large coefficients than higher-order terms.

- This may be incorporated by having a zero-mean Gaussian prior with large variances for low-order terms (indicating uncertainty) and small variances for high-order terms (i.e., you are certain that the coefficient is close to 0).

Alternative free energy decomposition

The “free energy” objective decomposes into prediction error and complexity:

$$\mathcal{F}[q] = \int_{\theta} q(\theta) \log \frac{1}{p(x|\theta)} d\theta + \int_{\theta} q(\theta) \log \frac{q(\theta)}{p(\theta)} d\theta$$

It can also be decomposed as an upper bound to negative model evidence:

$$\mathcal{F}[q] = \int_{\theta} q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta - \log p(x)$$

approximation to posterior (≥ 0)

model evidence

In this sense, a smaller free energy means 1) a better approximation of the posterior and/or 2) a better model for the given data.

Normalization

The solution for q^* led to a mysterious $1 / \exp$ term. Where does that come from?

It comes from the normalization constraint imposed on the Lagrangian.

If we plug the optimal form into the constraint function, we get:

$$\int \frac{1}{\exp(1 + \lambda)} p(\theta) d\theta - 1 = 0$$

Solving for λ gives:

$$\lambda = \log \int p(\theta) d\theta - 1$$

Mean-field

If there are multiple unknowns in the model, then you may choose to factorize q :

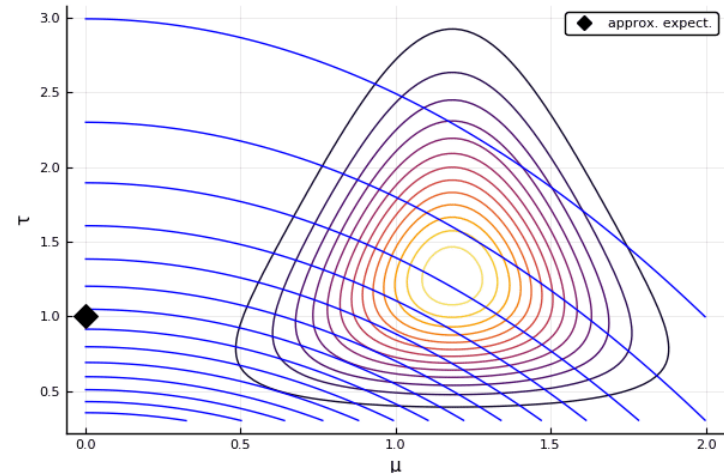
$$q(\theta, \sigma^2) := q(\theta)q(\sigma^2)$$

You would have multiple approximations, each dependent on the others.

-> Solutions must be iterated until convergence.

“Mean-field” is a common factorization choice, but may lead to poor performance.

“Structured” factorizations are richer, but require more manual derivation work.



Limitations

Common parametric distributions are not closed under nonlinear transformations.

- A squared Gaussian distributed random variable is not Gaussian distributed.

Typical simplifications of q are based on (in)dependence between variables.

- This may cause under-estimation of variance.

Not much is known about the stability of variational Bayesian estimators and some appear to be (at least numerically) unstable in practice.