Automatic relevance determination priors in Bayesian model selection: Application to nonlinear fluid-structure interaction systems

R. Sandhu[1], C. Pettit[2], M. Khalil[1,3], A. Sarkar[1] and D. Poirel[4]

[1]Carleton University, Ottawa, ON, Canada

[2]United States Naval Academy, Annapolis, MD, USA

[3]Sandia National Laboratories, Livermore, CA, USA

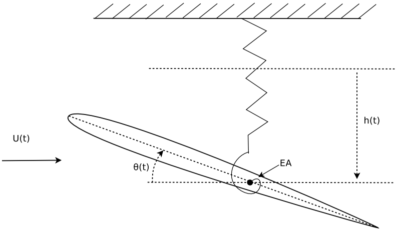
[4]Royal Military College of Canada, Kingston, ON, Canada

EMI 2018
Boston, MA, USA
May 29 - June 1, 2018

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.

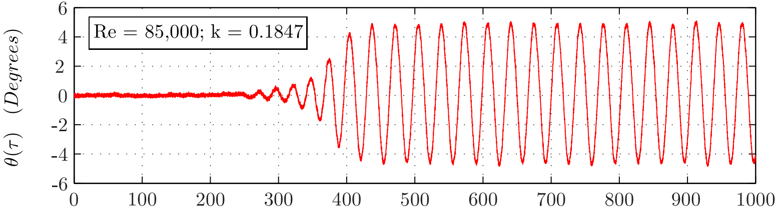
Outline

Problem definition



Nonlinear aeroelastic oscillator

- Pure pitch (dof: 1) limit cycle oscillations (LCO) of a 2-D rigid airfoil in a transitional Re regime



Goal: Identify the nature of unsteady and nonlinear aerodynamics causing the LCO by assimilating the noisy wind-tunnel observations.

Problem definition

Candidate model set: $\{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4\}$

Eq. of motion (Known):

$$I_{\mathrm{EA}}\ddot{\theta} + D\dot{\theta} + K\theta + K'\theta^3 = D'\mathrm{sign}(\dot{\theta}) + \frac{1}{2}\rho U^2 c^2 s\, C_M(\theta, \dot{\theta}, \ddot{\theta})$$

Possible models of aerodynamics (C_M):

$$\mathcal{M}_1 : \; Cm = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} + \sigma\xi(\tau)$$

$$\mathcal{M}_2 : \; C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} + e_5\theta^5 + e_6\theta^4\dot{\theta} + \sigma\xi(\tau)$$

$$\mathcal{M}_3 : \; \frac{\dot{C}_M}{B} + C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} + \frac{c_6}{B}\ddot{\theta} + \sigma\xi(\tau)$$

$$\mathcal{M}_4 : \; \frac{\dot{C}_M}{B} + C_M = e_1\theta + e_2\dot{\theta} + e_3\theta^3 + e_4\theta^2\dot{\theta} + e_5\theta^5 + e_6\theta^4\dot{\theta} + \frac{c_6}{B}\ddot{\theta} + \sigma\xi(\tau)$$

Measurement equation:

$$d_k = \theta_k + \epsilon_k \;\; ; \quad k = 1, \ldots, n_d \tag{1}$$

Bayesian model selection in discrete model space

Problem definition

Bayesian model selection in discrete model space

Given observational data $\mathbf{D} = \{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_{n_d}\}$ and a candidate model set $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2 \ldots \mathcal{M}_i \ldots \mathcal{M}_P\}$, the posterior model probability is calculated as

$$\mathrm{P}(\mathcal{M}_i|\mathbf{D}, \mathcal{M}) = \frac{\mathrm{p}(\mathbf{D}|\mathcal{M}_i)\mathrm{P}(\mathcal{M}_i|\mathcal{M})}{\mathrm{p}(\mathbf{D}|\mathcal{M})} \tag{2}$$

where $\mathrm{p}(\mathbf{D}|\mathcal{M}_i)$ is the model evidence, which embodies the principle of Ockham's razor,

$$\underbrace{\ln \mathrm{p}(\mathbf{D}|\mathcal{M}_i)}_{\text{Log-evidence}} = \int \mathrm{p}(\mathbf{D}|\phi)\mathrm{p}(\phi)d\phi = \underbrace{\mathrm{E}[\ln \mathrm{p}(\mathbf{D}|\phi, \mathcal{M}_i)]}_{\text{Goodness-of-fit}} - \underbrace{\mathrm{E}\left[\ln \frac{\mathrm{p}(\phi|\mathbf{D}, \mathcal{M}_i)}{\mathrm{p}(\phi|\mathcal{M}_i)}\right]}_{\text{Information gain (EIG)}} \tag{3}$$

Sandhu *et al.*, CMAME, 2017.
Sandhu *et al.*, JCP, 2016.
Sandhu *et al.*, CMAME, 2014.
Khalil *et al.*, JSV, 2013

Problem definition

Practical hurdles in implementing Bayesian model selection in discrete model space

- Sensitivity of parameter prior distribution to the posterior model probability or the model evidence
- Missing out on better candidate models

Solution: Automatic relevance determination (ARD)

Automatic Relevance Determination (ARD)

Reformulating the model selection problem:

- An encompassing model $\mathcal{M}$:

$$\frac{\dot{C}_M}{B} + C_M = a_1\theta + a_2\dot{\theta} + a_3\theta^3 + a_4\theta^2\dot{\theta} + a_5\theta^5 + a_6\theta^4\dot{\theta} + \frac{c_6}{B}\ddot{\theta} + \sigma\xi(\tau)$$

- The question we ask: Given measurements **D**, find the optimal model nested under the overly-complicated encompassing model?

Automatic Relevance Determination (ARD)

Physics-driven + Data-driven + Prior knowledge

Hybrid approach for assigning prior distributions by categorizing parameters based on prior knowledge about the aerodynamics as Required ($\phi_{\text{-}\psi}$) or Contentious (ϕ_{ψ})

$$\frac{\dot{C}_M}{B} + C_M = a_1\theta + a_2\dot{\theta} + a_3\theta^3 + a_4\theta^2\dot{\theta} + a_5\theta^5 + a_6\theta^4\dot{\theta} + \frac{c_6}{B}\ddot{\theta} + \sigma\xi(\tau)$$

Prior pdf, $p(\phi\|\boldsymbol{\psi}) = p(\phi_{\text{-}\psi})p(\phi_{\psi}\|\boldsymbol{\psi})$ Hyper-parameter, $\boldsymbol{\psi} = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$	
$p(\phi_{\text{-}\psi})$	$\mathcal{L}(B\|0.2, 50)\ \mathcal{U}(a_1\|-2, 0)\ \mathcal{U}(a_2\|-2, 0)\ \mathcal{L}(\sigma\|0.002, 50)$
$p(\phi_{\psi}\|\boldsymbol{\psi})$	ARD prior, $\mathcal{N}\left(a_3\|0, \frac{1}{\alpha_1}\right)\mathcal{N}\left(a_4\|0, \frac{1}{\alpha_2}\right)\mathcal{N}\left(a_5\|0, \frac{1}{\alpha_3}\right)\mathcal{N}\left(a_6\|0, \frac{1}{\alpha_4}\right)$

Automatic Relevance Determination (ARD)

Using hierarchical Bayes approach:

- Posterior pdf $\mathrm{p}(\boldsymbol{\psi}|\mathbf{d})$ of hyper-parameter vector $\boldsymbol{\psi}$,

$$\mathrm{p}(\boldsymbol{\psi}|\mathbf{d}) = \frac{\mathrm{p}(\mathbf{d}|\boldsymbol{\psi})\mathrm{p}(\boldsymbol{\psi})}{\mathrm{p}(\mathbf{d})} \tag{4}$$

- Assuming flat prior for $\mathrm{p}(\boldsymbol{\psi})$ Task: Stochastic optimization,

$$\boldsymbol{\psi}_{\mathrm{map}} = arg \max_{\boldsymbol{\psi}} \{\mathrm{p}(\mathbf{D}|\boldsymbol{\psi})\} \tag{5}$$

- Model evidence as a function of hyper-parameter Task: Evidence computation,

$$\mathrm{p}(\mathbf{D}|\boldsymbol{\psi}) = \int \mathrm{p}(\mathbf{D}|\phi)\mathrm{p}(\phi|\boldsymbol{\psi})d\phi \tag{6}$$

- Likelihood computation Task: State estimation,

$$\mathrm{p}(\mathbf{D}|\phi) = \prod_{k=1}^{n_d} \int \mathrm{p}(\mathbf{d}_k|\mathbf{u}_{j(k)}, \phi)\mathrm{p}(\mathbf{u}_{j(k)}|\mathbf{d}_{1:k-1}, \phi)d\mathbf{u}_{j(k)} \tag{7}$$

Automatic Relevance Determination (ARD)

Numerical implementation

- Evidence optimization: Derivative-free methods including line-search, pattern search, simplex method, evolutionary algorithms; and many others.
- Evidence computation: Chib-Jeliazkov method, Transitional MCMC, Power posteriors, Nested sampling, Annealed importance sampling, Harmonic mean estimator, Gauss-Hermite quadrature; and many others.
- MCMC sampler for Chib-Jeliazkov method: Metropolis-Hastings, Gibbs, TMCMC, adaptive Metropolis, Delayed Rejection Adaptive Metropolis(DRAM); and many others
- State estimation: Kalman filter, Extended Kalman filter, unscented Kalman filter, ensemble Kalman filter, particle filter; and many others.

Numerical results: Unidimensional ARD

ARD prior for a relevant parameter

Model proposed same as the data-generating model:

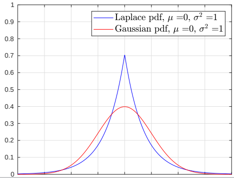
$$\frac{\dot{C}_M}{B} + C_M = a_1\theta + a_2\dot{\theta} + a_3\theta^3 + a_4\theta^2\dot{\theta} + \frac{c_6}{B}\ddot{\theta} + \sigma\xi(\tau), \tag{8}$$

Case 1: Gaussian ARD prior:

$$p(\phi|\psi) = \mathcal{L}(B|0.2, 50)\mathcal{U}(a_1|-2, 0)\mathcal{U}(a_2|-2, 0)\ \mathcal{N}(a_3|0, 1/\alpha)\ \mathcal{U}(a_4|-600, 0)\mathcal{L}(\sigma|0.002, 50) \tag{9}$$

Case 2: Laplace ARD prior:

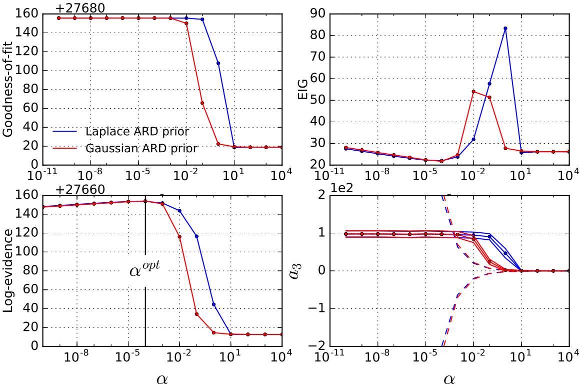
$$p(\phi|\psi) = \mathcal{L}(B|0.2, 50)\mathcal{U}(a_1|-2, 0)\mathcal{U}(a_2|-2, 0)\ \mathcal{LP}(a_3|0, 1/\alpha)\ \mathcal{U}(a_4|-600, 0)\mathcal{L}(\sigma|0.002, 50) \tag{10}$$



Numerical results: Unidimensional ARD

Observations:

- Change in log-evidence driven by loss of goodness-of-fit due to removal of a_3
- Log-evidence has higher slope near maxima and is minimally sloped elsewhere
- Both Laplace prior and Gaussian prior results in same parameter sparsity level.



Numerical results: Unidimensional ARD

Effect of using zero-mean ARD priors on parameter estimates

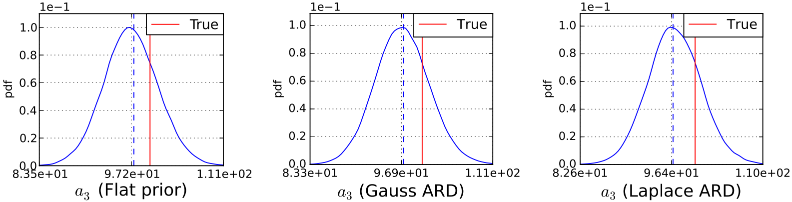


Figure: Comparison of marginal posterior pdf of parameter a_3 obtained using optimized Gaussian and Laplace ARD prior, compared with the marginal posterior obtained using a flat prior for parameter a_3.

Numerical results: Unidimensional ARD

ARD prior for an irrelevant parameter

Model proposed has an additional term than the data-generating model:

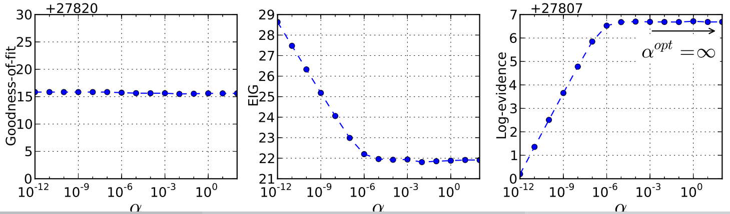
$$\frac{\dot{C}_M}{B} + C_M = a_1\theta + a_2\dot{\theta} + a_3\theta^3 + a_4\theta^2\dot{\theta} + {\color{red}a_5\theta^5} + \frac{c_6}{B}\ddot{\theta} + \sigma\xi(\tau), \tag{11}$$

$$\begin{aligned} \mathrm{p}(\phi|\psi) =& \mathcal{L}(B|0.2, 50)\mathcal{U}(a_1|-2, 0)\mathcal{U}(a_2|-2, 0)\mathcal{U}(a_3|-250, 250) \\ & \mathcal{U}(a_4|-600, 0){\color{red}\mathcal{N}(a_5|0, 1/\alpha)}\mathcal{L}(\sigma|0.002, 50) \end{aligned} \tag{12}$$

Observations:

- The change in log-evidence is driven by the decrease in Complexity (EIG) due to the removal of irrelevant parameter.

- Log-evidence is flat in regions higher the optimal hyperparameter



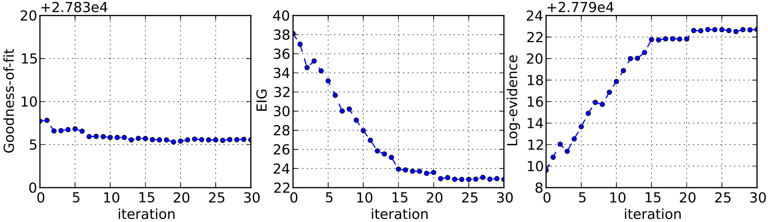
Numerical results: Multidimensional ARD

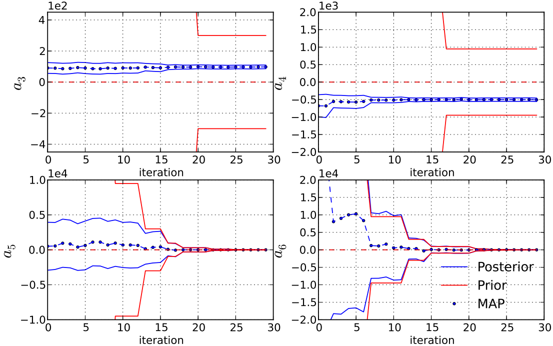
Model proposed:

$$\frac{\dot{C}_M}{B} + C_M = a_1\theta + a_2\dot{\theta} + a_3\theta^3 + a_4\theta^2\dot{\theta} + a_5\theta^5 + a_6\theta^4\dot{\theta} + \frac{c_6}{B}\ddot{\theta} + \sigma\xi(\tau) \tag{13}$$

Prior pdf, $p(\phi\|\boldsymbol{\psi}) = p(\phi_{\text{-}\psi})p(\phi_{\psi}\|\boldsymbol{\psi})$ Hyper-parameter, $\boldsymbol{\psi} = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$	
$p(\phi_{\text{-}\psi})$	$\mathcal{L}(B\|0.2, 50)\,\mathcal{U}(a_1\|-2, 0)\,\mathcal{U}(a_2\|-2, 0)\,\mathcal{L}(\sigma\|0.002, 50)$
$p(\phi_{\psi}\|\boldsymbol{\psi})$	ARD prior, $\mathcal{N}\left(a_3\|0, \frac{1}{\alpha_1}\right)\mathcal{N}\left(a_4\|0, \frac{1}{\alpha_2}\right)\mathcal{N}\left(a_5\|0, \frac{1}{\alpha_3}\right)\mathcal{N}\left(a_6\|0, \frac{1}{\alpha_4}\right)$

Numerical results: Multidimensional ARD





Numerical results: Multidimensional ARD

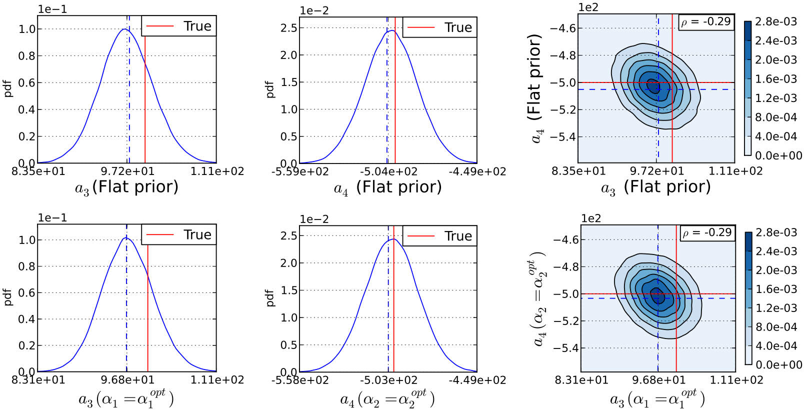


Figure: Comparison of marginal and joint posterior pdf of relevant parameters a_3 and a_4 for ARD prior with optimal hyper-parameters and flat priors pdf.

Conclusion

Conclusion

- The concept of automatic relevance determination (ARD) is exploited as an automatic model selection tool with application to nonlinear dynamical systems modelled using stochastic ordinary differential equations (ODE).
- ARD approach is validated using a synthetically generated nonlinear aeroelastic oscillations.
- Both Laplace and Gaussian ARD prior produced same parameter sparsity level.
- Derivative-free optimization techniques with bound constraint are well-suited for optimizing model evidence due to the flatness of objective function (Log-evidence) away from maxima.

Future direction

- Using gradient/hessian information to expedite the optimization of model evidence.
- Comparing the ARD approach to LASSO/Ridge regression techniques.

R. Sandhu, C. Pettit, M. Khalil, D. Poirel, A. Sarkar, Bayesian model selection using automatic relevance determination for nonlinear dynamical systems, Computer Methods in Applied Mechanics and Engineering (2017).

Acknowledgement

Thank you.

Financial contributions

- Department of National Defence, Canada
- Ontario Graduate Scholarship program
- Carleton University, Ottawa, Canada
- Natural Sciences and Engineering Research Council, Canada

Supercomputers used

- HP Linux Cluster, Carleton University, Ottawa, Canada
- CLUMEQ, McGill University, Montreal, Canada
- SciNet, University Of Toronto, Toronto, Canada

External libraries used

- Dakota, UQ Toolkit (UQTk) [Developed by Sandia National Lab]
- Armadillo (Linear algebra library for C++)