Constrained parameter estimation of δ-domain models

February 27, 2020

This report demonstrates the performance of the δ-domain identification framework on a simulation example.

1 Model structure identification

Equivalence between the lag models and δ-domain models was established in [1]. For a input-output model with a defined lags

$$\boldsymbol{y}(t) = f\big(\{y(t-k)\}_{k=1}^{n_y}, \{u(t-k+n_y+1)\}_{k=n_y+1}^{n_y+n_u}\big), \tag{1}$$

there can be established an equivalent representation with δ-operator, where the output of the NARX model is the highest order derivative of the registered output, $\delta^n y(t)$, and input vector takes the following form:

$$\boldsymbol{x}(t) = \left[\begin{array}{cccccccc} \delta^{n-1}y(t) & \dots & \delta y(t) & y(t) & \delta^{n-1}u(t-1) & \dots & \delta u(t) & u(t) \end{array} \right]^\top. \tag{2}$$

The unknown model is approximated with a sum of polynomial basis functions up to second degree ($\lambda = 3$), rendering the following structure

$$\boldsymbol{y}(t) = \theta^0 + \sum_{i=1}^{d} \theta_i x_i(t) + \sum_{i=1}^{d}\sum_{j=i}^{d} \theta_{i,j} x_i(t) x_j(t) + \sum_{i=1}^{d}\sum_{j=i}^{d}\sum_{k=j}^{d} \theta_{i,j} x_i(t) x_j(t) x_k(t) + e(t). \tag{3}$$

The performance of δ-domain identification framework is tested on a simulated data for Van-der-Pol oscillator (VDPO) with varying damping strength. The dynamics of VDPO is described by a non-linear second-order ODE:

$$\frac{\delta^2}{\delta t^2} y(t) = \mu(1 - y^2(t)) \frac{\delta}{\delta t} y(t) - y(t) + u(t), \tag{4}$$

where $u(t)$ is an excitation signal (in this case, sum of sinusoids). The damping coefficient μ was assigned the following values for the MC simulations

$$\mu = \left[\begin{array}{ccccccc} 0.0625 & 0.125 & 0.25 & 0.3 & 0.5 & 0.8 & 1. \end{array} \right]$$

The number and order of significant terms are identified within the EFOR-CMSS truncated using Bayesian information criteria

The significant terms identified by the algorithm are presented in Table **??**

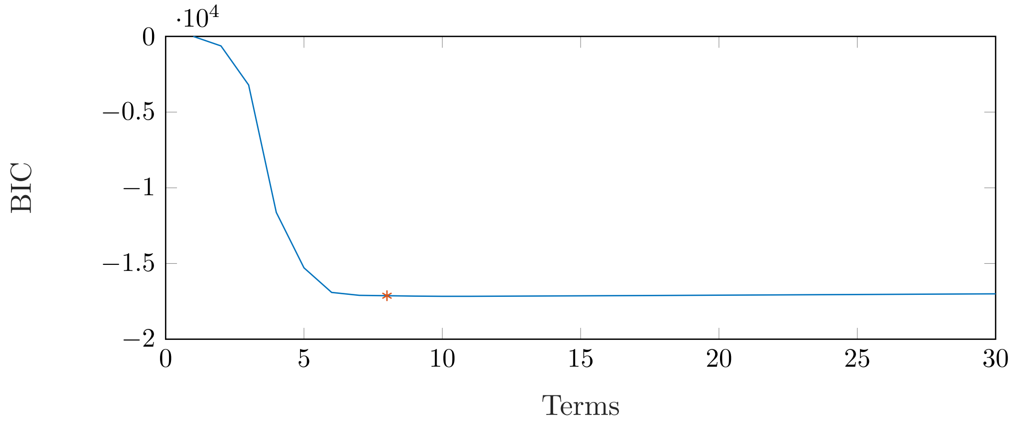

Figure 1: Evolution of BIC with growing number of parameters in the model

Table 1: Significant terms and corresponding coefficients identified in EFOR-CMSS algorithm.

Step	Terms	V1	V2	V3	V4	V5	V6	V7	AERR(%)	BIC
1	$y(t)$	-102.95	-105.04	-112.39	-117.57	-132.86	-151.46	-160.09	58.525	0
2	$u(t)$	95.95	96.75	96.52	96.85	96.29	96.28	97.22	13.25	-627.5513
3	$y(t)y(t)\delta^1y(t)$	-0.6	-1.19	-2.39	-2.87	-4.8	-7.76	-9.72	13.368	-3208.4451
4	$\delta^1u(t)$	0.93	0.94	0.93	0.93	0.91	0.89	0.89	9.483	-11627.532
5	$\delta^1y(t)$	-0.42	0.14	1.26	1.7	3.41	5.72	7.16	3.694	-15285.6711
6	$y(t)y(t)y(t)$	0.71	1.23	3.17	4.67	9.33	14.11	16.79	0.63	-16910.4537
7	$y(t)\delta^1y(t)u(t)$	-0.02	-0.06	-0.14	-0.17	-0.25	-0.57	-0.97	0.075	-17102.6631
8	$y(t)y(t)u(t)$	-0.26	-0.46	-0.84	-0.68	-1.42	-1.96	-1.32	0.007	-17127.8805

2 Direct estimation of external model parameters

In order to link the external and internal parameters, an arbitrary polynomial function is formed from either a single parameter vector or a pair of vectors. The number of unknown parameters is defined by the number of avaliable datasets. This section demonstrates the direct estimation procedure and the justification for selecting polynomial terms for curve fitting.

Model structure for K datasets:

$$\underbrace{\bar{\mathbf{Y}}}_{\mathrm{T}\,K\times 1} = \underbrace{\bar{\Phi}}_{\mathrm{T}\,K\times NK}\underbrace{\bar{\Theta}}_{NK\times 1}, \tag{5}$$

where the block matrices have the following structure:

$$\underbrace{\bar{\mathbf{Y}}}_{\mathrm{T}\,K\times 1} = \begin{bmatrix} \underbrace{\boldsymbol{y}^1}_{\mathrm{T}\times 1} \\ \underbrace{\boldsymbol{y}^2}_{\mathrm{T}\times 1} \\ \vdots \\ \underbrace{\boldsymbol{y}^K}_{\mathrm{T}\times 1} \end{bmatrix}; \quad \underbrace{\bar{\Phi}}_{\mathrm{T}\,K\times NK} = \begin{bmatrix} \underbrace{\Phi^1}_{\mathrm{T}\times N} & \dots & \dots & \mathbb{O} \\ \mathbb{O} & \underbrace{\Phi^2}_{\mathrm{T}\times N} & \dots & \mathbb{O} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbb{O} & \dots & \dots & \underbrace{\Phi^K}_{\mathrm{T}\times N} \end{bmatrix}; \quad \underbrace{\bar{\Theta}}_{NK\times 1} = \begin{bmatrix} \underbrace{\theta^1}_{N\times 1} \\ \underbrace{\theta^2}_{N\times 1} \\ \vdots \\ \underbrace{\theta^K}_{N\times 1} \end{bmatrix}. \tag{6}$$

The relationship of the design parameters known from the experiments and the internal parameters of NARMAX model is defined by the following linear function:

$$\underbrace{\Theta}_{N\times K} = \underbrace{B}_{N\times L}\underbrace{A}_{L\times K}, \tag{7}$$

where A is the matrix where each row is a function of the vector of design parameters. The example structure is

$$A = \left[\begin{array}{cccccc} \mathbb{I}_{K\times 1} & L_{K\times 1} & D_{K\times 1} & LD_{K\times 1} & L^2_{K\times 1} & D^2_{K\times 1} \end{array}\right]^\top, \tag{8}$$

and where B denotes the matrix of unknown coefficients of a hypersurface of order L that maps a point in external parameter space, $\xi^k = (L_k, D_k)$, onto the point in the space of internal parameters, θ^k. In can be seen that $\bar{\Theta} = \text{vec}(\Theta)$, then

$$\underbrace{\bar{\Theta}}_{NK\times 1} = \text{vec}\left(\underbrace{B}_{N\times L}\underbrace{A}_{L\times K}\right). \tag{9}$$

This vectorisation can be obtained using Kronecker product:

$$\text{vec}\left(\underbrace{B}_{N\times L}\underbrace{A}_{L\times K}\right) = (\underbrace{A^\top}_{K\times L} \otimes \underbrace{\mathbb{I}}_{N\times N})\underbrace{\text{vec}(B)}_{NL\times 1}. \tag{10}$$

Denoting the result of Kronecker product as $\underbrace{\mathbf{Kr}}_{NK\times NL} \triangleq (\underbrace{A^\top}_{K\times L} \otimes \underbrace{\mathbb{I}}_{N\times N})$ and vectorised coefficient matrix as $\underbrace{\bar{\mathbf{B}}}_{NL\times 1} \triangleq \text{vec}(B)$ yields the following:

$$\underbrace{\bar{\Theta}}_{NK\times 1} = \underbrace{\mathbf{Kr}}_{KN\times NL}\underbrace{\bar{\mathbf{B}}}_{NL\times 1}. \tag{11}$$

Substituting the above expression into (5) renders an expression that directly links the design parameters and the timeseries data

$$\underbrace{\bar{\mathbf{Y}}}_{\mathrm{T}K\times 1} = \underbrace{\bar{\Phi}}_{\mathrm{T}K\times NK}\underbrace{\mathbf{Kr}}_{KN\times NL}\underbrace{\bar{\mathbf{B}}}_{NL\times 1}, \tag{12}$$

where $\bar{\mathbf{B}}$ is the unknown vector and all other factors are known form the experiments or defined prior to structure identification.

The representation (12) allows estimating the coefficients in $\bar{\mathbf{B}}$ directly from the timeseries data bypassing the intermediate estimation of the internal coefficients in NARX model.

The following condition must be satisfied:

$$\text{rk}(\bar{\Phi}\mathbf{Kr}) \geq NL. \tag{13}$$

The rank of the linear system (12) satisfies the following:

$$\text{rk}(\bar{\Phi}\mathbf{Kr}) \leq \min\left(\text{rk}(\bar{\Phi}), \text{rk}(\mathbf{Kr})\right), \tag{14}$$

where the rank of Kronecker product can be found as

$$\text{rk}(\mathbf{Kr}) = \text{rk}(\underbrace{A^\top}_{K\times L})\,\text{rk}(\underbrace{\mathbb{I}}_{N\times N}), \tag{15}$$

thus the matrix A composed of by-element combinations of external parameter vector(s) must be of rank K.

3 Constrained estimation

A common treatment for models that suffer from bad generalisation is to introduce a constrained LS problem, where the constraint is normally posed on the unknown parameter vector:

$$\hat{\beta}^i = \arg\min \left\| \left(\bar{\theta}^i - X\beta^i \right) \right\|^2, \quad f_R(\beta^i) < \gamma. \tag{16}$$

where λ is a pre-specified parameter that defines the size of the constraint in the parameter space Ξ. Lagrangian formulation of the constrained problem is called regularised Least Squares (RLS),

$$\beta^i = \arg\min \left\{ \left\| \left(\bar{\theta}^i - X\beta^i \right) \right\|^2 + \lambda f_R(\beta^i) \right\} \tag{17}$$

where the regularisation coefficient λ is directly linked to γ in (16). Different types of the constraint function are be considered depending on the problem.

3.1 Tikhonov regularisation

Innovation regularisation constrains the 2-norm of the parameter vector

$$f_R(\beta^i) = \left\| \beta^i \right\|_2^2. \tag{18}$$

This is the only RLS formulation that has a closed form solution that is usually obtained for the normalised data.

$$\hat{\beta}^i_{RLS} = (\mathbf{R}^*_{aa} + \lambda\, \mathbb{I}_M)^{-1} (\mathbf{A}^\star)^\top \bar{\theta}^i, \tag{19}$$

This solution is referred to as ridge regression, because increasing λ shrinks the coefficients β_j. The shrinkage is shown is best interpreted via singular values decomposition (SVD) of the normalised data matrix. Denote the SVD of $\mathbf{A}^\star$ as

$$\underbrace{\mathbf{A}^\star}_{K\times M} = \underbrace{\mathbf{U}}_{K\times M} \underbrace{\mathbf{D}}_{M\times M} \underbrace{\mathbf{V}^\top}_{M\times M}, \tag{20}$$

where columns of $\mathbf{U}$ are principal components of $\mathbf{A}^\star$, diagonal elements of $\mathbf{D}$ are the singular values, and where $\mathbf{V}$ is the rotation. All orthonormality assumption are the same as in the general case. The ridge regression (19) then takes form

$$\hat{\beta}^i_{RLS} = \left(\mathbf{V}\mathbf{D}\mathbf{U}^\top \mathbf{U}\mathbf{D}\mathbf{V}^\top + \lambda\, \mathbb{I}_M \right)^{-1} \left(\mathbf{U}\mathbf{D}\mathbf{V}^\top \right)^\top \bar{\theta}^i. \tag{21}$$

Simple linear algebra yields the following

$$\hat{\beta}^i_{RLS} = \mathbf{V} \left(\mathbf{D}^2 + \lambda\, \mathbb{I}_M \right)^{-1} \mathbf{V}^\top \mathbf{V}\mathbf{D}\mathbf{U}^\top \bar{\theta}^i. \tag{22}$$

The finale expression is

$$\hat{\beta}^i_{RLS} = \mathbf{V} \left(\mathbf{D}^2 + \lambda\, \mathbb{I}_M \right)^{-1} \mathbf{D}\mathbf{U}^\top \bar{\theta}^i, \tag{23}$$

which can be compared to the SVD of OLS regression:

$$\hat{\beta}^i_{OLS} = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}^\top \bar{\theta}^i \qquad \left(\approx (\mathbf{A}^\star)^{-1} \bar{\theta}^i \right). \tag{24}$$

It can be seen from (23)-(24) that in Tikhonov regularisation the inverse of the diagonal matrix is obtained as $\mathbf{D}/(\mathbf{D}^2 + \lambda\,\mathbb{I}_M)$ where λ is non-negative. Increasing regularisation coefficient thus leads to shrinkage of the singular values, and the estimates asymptotically approach zero. This shows that in Tikhonov regularisation λ quantifies the trade-off between the bias and the variance in the estimates. Small regularisation coefficient leads to near-OLS solution that overfits the model to the training data, while large value leads to biased estimates and drives all coefficients to near-zero values.

3.2 LASSO regularisation

While in the ridge regression LSEs asymptotically approach zero, none of the parameters can be zeroed-out explicitly if the model structure is overly detailed. Another formulation of RLS, called Least absolute shrinkage and selection operator (LASSO) regression, performs variable selection and the regularisation simultaneously by imposing the l_1 penalty:

$$f_R(\beta^i) = \left\|\beta^i\right\|_1. \tag{25}$$

The Lagrangian optimisation problem then takes for of basis pursuit de-noising that can be solved numerically using quadratic programming or convex techniques. This report uses the shooting algorithm proposed in [2] because of its relative simplicity. The results of ridge estimation are used as the initial point in convex optimisation searching for the LASSO solution.

3.3 Selection of the regularisation coefficient

The important stage of solving RLS problem is selecting the regularisation parameter that will result into a interpretable but parsimonious model. The constraint γ in (**??**) is often selected arbitrary since the shape of the parameter space is unknown. As a result, finding λ relies on iterative schemes most of which do not guarantee convergence [CITE MANY]. For the lack of universal approach for selecting the optimal value of λ, the choice of the method remains application-specific.

In this report, ridge estimation is applied to a fixed model structure, hence Aikaike's information criterion (AIC) may be used

$$\mathrm{AIC}_\lambda = 2p - 2\log p(\bar{\theta}^i \mid \beta^i, \lambda), \tag{26}$$

where the first term quantifies model complexity and the second term is the log-likelihood of the selected model fitting the data. The model complexity is determined as simply the trace of the hat matrix of the ridge estimator

$$p = \mathrm{tr}(\mathbf{H}_{RLS}) = \mathrm{tr}((\mathbf{A}^\top)(\mathbf{R}_{aa} + \lambda\,\mathbb{I}_M)^{-1}\mathbf{A}^\top) \tag{27}$$

. Bayesian information criterion (BIC), assigns a larger penalty to the model complexity

$$\mathrm{BIC}_\lambda = 2\log(n)p - 2\log p(\bar{\theta}^i \mid \beta^i, \lambda), \tag{28}$$

where n is the number of data points used for parameter estimation. Both AIC and BIC are aim to estimate

When it comes to finding a parsimonious model, it may be more reasonable to access model's prediction performance instead of explicitly penalising its complexity. Cross-validation procedure This work

Selcting regularisation coefficient for LASSO regression is more nuanced as different model structure may arise for different values of λ. The most popular approach described in the literature uses cross-validation. Both the data matrix and the response vector are partitioned into pairs. Then each pair is exuded from the

4 Results

The estimated coefficients are presented in Table 2, and the surface fitting results for each internal parameter are illustrated in Figure **??**

Table 2: Polynomial coefficients estimated via ordinary LS.

Step	Terms	β_0	β_1	β_2	β_3
1	$y(t)$	-99.66	-36.65	-86.47	62.92
2	$u(t)$	95.83	7.52	-19.58	13.48
3	$y(t)y(t)\delta^1 y(t)$	-0.01	-9.42	-0.47	0.16
4	$\delta^1 u(t)$	0.92	0.13	-0.48	0.32
5	$\delta^1 y(t)$	-1.02	9.5	-1.36	0.03
6	$y(t)y(t)y(t)$	-0.29	10.27	24.3	-17.61
7	$y(t)\delta^1 y(t)u(t)$	0.04	-1.04	1.69	-1.66
8	$y(t)y(t)u(t)$	-0.43	1.17	-10.6	8.54

Table 3: Polynomial coefficients estimated via Tikhonov regularisation.

Step	Terms	β_0	β_1	β_2	β_3
1	$y(t)$	-99.66	-36.65	-86.47	62.92
2	$u(t)$	95.83	7.52	-19.58	13.48
3	$y(t)y(t)\delta^1 y(t)$	-0.01	-9.42	-0.47	0.16
4	$\delta^1 u(t)$	0.92	0.13	-0.48	0.32
5	$\delta^1 y(t)$	-1.02	9.5	-1.36	0.03
6	$y(t)y(t)y(t)$	-0.29	10.27	24.3	-17.61
7	$y(t)\delta^1 y(t)u(t)$	0.04	-1.04	1.69	-1.66
8	$y(t)y(t)u(t)$	-0.43	1.17	-10.6	8.54

Table 4: Polynomial coefficients estimated via LASSO regularisation.

Step	Terms	β_0	β_1	β_2	β_3
1	$y(t)$	-99.66	-36.65	-86.47	62.92
2	$u(t)$	95.83	7.52	-19.58	13.48
3	$y(t)y(t)\delta^1 y(t)$	-0.01	-9.42	-0.47	0.16
4	$\delta^1 u(t)$	0.92	0.13	-0.48	0.32
5	$\delta^1 y(t)$	-1.02	9.5	-1.36	0.03
6	$y(t)y(t)y(t)$	-0.29	10.27	24.3	-17.61
7	$y(t)\delta^1 y(t)u(t)$	0.04	-1.04	1.69	-1.66
8	$y(t)y(t)u(t)$	-0.43	1.17	-10.6	8.54