Regularised Least Squares estimation of external parameters

November 29, 2019

Preliminaries

Recall the notation from the previous reports:

$k = 1, \ldots, K$ - index of the dataset used for identification.

$t = 1, \ldots, T$ - discrete time index.

N_s - number of significant terms in the model.

$i = 1, \ldots, N_s$ - term index.

$\Xi = \mathbb{R}^2$ - space of external parameters.

$\Theta = \mathbb{R}^{N_s}$ - space of internal parameters.

θ_i^k - an internal regression parameter corresponding to the i-th term in the k-th dataset.

$\theta^k = \{\theta_i^k\}_{i=1}^{N_s}$ - vector of internal parameters.

The analysis deals with the time-series data from auxetic foam vibration tests introduced in the previous report. A non-linear FIR filter is considered with the input lag of 3. The filter structure is approximated using polynomial regressors of order 2. Number the significant terms and corresponding scaling parameters θ are estimated via EFOR-CMSS algorithm based on 8 datasets out of 10 available. The remaining to datasets (C3 and C8) are used for model validation.

This report deals only with identification of the relationship between the estimated internal parameters θ^k and the pre-specified design settings $\xi^k = (L_{cut}, D_{rlx})$. The first section examines the effect on the polynomial structure and sampling from the small subspace of external parameters on the generalisation of the identified model. The remaining sections briefly describe approaches to improving the estimation procedure. The results section only demonstrates numerical results and does not draw any conclusions as this report is intermediate.

1 Collinearity in the selected model

Recall the assumed polynomial model that links internal parameters to the external ones:

$$\theta_i^k(L_{cut}, D_{rlx}) = \beta_0 + \beta_1 L_{cut} + \beta_2 D_{rlx} + \beta_3 L_{cut}^2 + \beta_4 D_{rlx}^2 + \beta_5 L_{cut} D_{rlx}, \qquad i = 1, \ldots, N_s, \quad (1)$$

where polynomial coefficients $\beta^i = [\beta_0 \ldots \beta_5]$ are unknown. In the original paper [1] they are estimated via batch LS estimator as follows. First, the matrix of "independent" variables is formed from the external parameter values:

$$\mathbf{A} = \begin{bmatrix} 1 & L & D & L \times L & D \times D & L \times D \end{bmatrix}, \quad (2)$$

where $L = \{L_{cut}^k\}_{k=1}^K$, $D = \{D_{rlx}^k\}_{k=1}^K$, the number of columns $M + 1$, where M is the order of polynomial (1), and where $\times$ denotes the by-element product such that

$$L \times L = \left[L_{cut}^{(k)} L_{cut}^{(k)}\right]_{k=1}^K.$$

Then, for each internal parameter a vector of estimated values is formed across K training datasets:

$$\bar{\theta}^i = \left\{\theta_i^1, \theta_i^1, \ldots, \theta_i^K\right\}^\top.$$

Note that while the internal parameters are considered to be fixed in the estimation, their estimates are obtained from the random data and therefore are random variables

$$\bar{\theta}^i \sim \mathcal{N}(\mu_i, \sigma_i^2).$$

Remark 1.1: *For the purposes of this analysis the factors contributing to this variance are not considered and the vector of LSEs is treated as a response signal in the LS formulation.*

Estimation of the polynomial coefficients for the i-th internal parameter β^i is an ordinary least squares (OLS) problem:

$$\hat{\beta}^i = \underset{\beta \in \mathbb{R}}{\arg\min} \left\{ \left\| \left(\bar{\theta}^i - \mathbf{A}\beta^i \right) \right\|_2^2 \right\} \quad (3)$$

with closed form solution

$$\hat{\beta}_{OLS}^i = \mathbf{R}_{aa}^{-1} \mathbf{A}^\top \bar{\theta}^i, \quad (4)$$

where $\mathbf{R}_{aa}^* = (\mathbf{A}^*)^\top \mathbf{A}^*$ is the regression matrix. The OLS estimator is unbiased and provides the solution that best fits the training data, where the fit is characterised by the sum of squared residuals (RSS). The unbiasedness is usually at the cost of the estimate variance and thus does not guarantee good generalisation of the model with estimated parameters.

It can be seen from Table 1 in previous report that the experimental data is collected from the narrow subspace of design parameters. Combined with the polynomial structure that means that the columns of the data matrix $\mathbf{A}$ are powers of one another, this may lead to severe collinearity problem. The quickest way to access collinearity is to check the condition number of the scaled matrix $\mathbf{A}^*$. As

per recommendation in [cite Seber and Lee], the data is not centred to see the effect on the raw data perturbation on the LSE variance:

$$a^*_{k,j} = \frac{a_{k,j}}{\left\{ \sum_{k=1}^{K} a^2_{k,j} \right\}^{1/2}},$$

where j is the column index. For the given data $\kappa(\mathbf{A}^*) = 1083.57$ which indicates that some eigenvalues of the regression matrix $\mathbf{R}^*_{aa} = (\mathbf{A}^*)^\top \mathbf{A}^*$ are close to zero:

$$\text{eig}(\mathbf{R}^*_{aa}) = \begin{bmatrix} 4.78 \times 10^{-6} \\ 3.50 \times 10^{-5} \\ 0.0015 \\ 0.0199 \\ 0.3625 \\ 5.6161 \end{bmatrix}$$

Some collinearity metrics for centred and scaled matrix $\mathbf{A}^\star_s$, where $\mathbf{A} = [] 1, \mathbf{A}_s$ are summarised in Table 1. The correlation of each columns and linear combinations of other columns is assessed by computing the coefficients of determination for individual columns

$$R^2_j = 1 - \frac{(a^\star_j)^\top H_j a^\star_j}{\sum_i (a^\star_j - \hat{a}^\star_j)^2},$$

where $\mathbf{H}_j = \mathbf{A}^\star_j((\mathbf{A}^\star_j)^\top \mathbf{A}^\star_j)^{-1}(\mathbf{A}^\star_j)^\top$ is the hat matrix and where $\mathbf{A}^\star_j$ is the data matrix that excludes the column $a^\star_j$. The ratio in this case quantifies the distance between the column of interest and its linear regression on other columns. The smaller this distance is, the higher determination coefficient for the column $a^\star_j$ will be. However, coefficients of determination do not indicate which particular columns have near-linear dependence. Thus, the effect of collinearity on LSE variance is assessed via variance inflation coefficients

$$\text{VIF}_j = \frac{\text{Var}(\beta_j)}{\sigma^2_j} = \frac{\sigma^2_j}{1 - R^2_j},$$

while the number of correlated column maybe better assessed from eigenvalues of the scaled and centred regression matrix are

$$\text{eig}(\mathbf{R}^\star_{aa}) = \begin{bmatrix} 0.0002 \\ 0.0004 \\ 0.0075 \\ 1.9647 \\ 3.0271 \\ 8 \end{bmatrix}$$

It can be seen that at least two columns of the normalised data matrix suffer from collinearity, and its effect on variance inflation confirms that the model may suffer bad generalisation outside of the considered subspace of design parameters ξ. This indicates the need for regularisation of the estimation procedure in order to reduce the dispersion of the parameter LSEs.

Table 1: Collinearity metrics of the data.

Column	L	R	$L \times R$	$L \times L$	$R \times R$
R_j^2	0.9992	0.9988	0.9997	0.9992	0.9985
VIF_j	1256.04	862.29	2867.46	1212.49	665.06

2 Regularised Least Squares

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

$$\hat{\beta}^i = \underset{\beta \in \mathbb{R}}{\arg\min} \left\| \left(\bar{\theta}^i - \mathbf{A}\beta^i \right) \right\|_2^2, \quad f_R(\beta^i) < \gamma. \tag{5}$$

where γ 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),

$$\hat{\beta}^i = \underset{\beta \in \mathbb{R}}{\arg\min} \left\{ \left\| \left(\bar{\theta}^i - \mathbf{A}\beta^i \right) \right\|_2^2 + \lambda f_R(\beta^i) \right\} \tag{6}$$

where the regularisation coefficient λ is directly linked to γ in (5) CITE. Different types of the constraint function can be considered depending on the problem.

2.1 Tikhonov regularisation

Innovation regularisation constrains the 2-norm of the parameter vector

$$f_R(\beta^i) = \left\| \beta^i \right\|_2^2. \tag{7}$$

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

$$\hat{\beta}^i_{RLS} = (\mathbf{R}^*_{aa} + \lambda \, \mathbb{I}_M)^{-1} (\mathbf{A}^\star)^\top \bar{\theta}^i, \tag{8}$$

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

$$\underbrace{\mathbf{A}^\star}_{M \times K} = \underbrace{\mathbf{U}}_{M \times K} \underbrace{\mathbf{D}}_{K \times K} \underbrace{\mathbf{V}^\top}_{K \times K}, \tag{9}$$

where columns of $\mathbf{U}$ are principal components of $\mathbf{A}^\star$, diagonal elements of $\mathbf{D}$, and where $\mathbf{V}$ is the rotation. All orthonormality assumption are the same as in the general case. The ridge regression (8) then takes form

$$\hat{\beta}^i_{RLS} = \left(\mathbf{V}\mathbf{D}\mathbf{U}^\top \mathbf{U}\mathbf{D}\mathbf{V}^\top + \lambda \, \mathbb{I}_M \right)^{-1} \left(\mathbf{U}\mathbf{D}\mathbf{V}^\top \right)^\top \bar{\theta}^i. \tag{10}$$

Simple linear algebra yields the following

$$\hat{\beta}^i_{RLS} = \mathbf{V} \left(\mathbf{D}^2 + \lambda \, \mathbb{I}_M \right)^{-1} \mathbf{V}^\top \mathbf{V}\mathbf{D}\mathbf{U}^\top \bar{\theta}^i. \tag{11}$$

The finale expression is

$$\hat{\beta}^i_{RLS} = \mathbf{V}\left(\mathbf{D}^2 + \lambda\,\mathbb{I}_M\right)^{-1}\mathbf{D}\mathbf{U}^\top\bar{\theta}^i, \tag{12}$$

which can be compared to the SVD of OLS regression:

$$\hat{\beta}^i_{OLS} = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}^\top\bar{\theta}^i, \qquad \left(\approx (\mathbf{A}^\star)^{-1}\bar{\theta}^i\right). \tag{13}$$

It can be seen from (12)-(13) 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 asymptotic shrinkage of the estimates to 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.

2.2 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. The regularising term is the 1-norm of the parameter vector:

$$f_R(\beta^i) = \left\|\beta^i\right\|_1. \tag{14}$$

zeroing out schemes.

Remark 2.1: *Results of ridge estimation can be used as the initial point in convex optimisation while solving the LASSO regression.*

2.3 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 (5) 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:

$$\mathrm{AIC}_\lambda = p + \frac{1}{2}\log p(\bar{\theta}^i \mid \beta^i, \lambda), \tag{15}$$

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

$$p = \mathrm{tr}(\mathbf{H}_{RLS}) = \mathrm{tr}((\mathbf{A}^\top)(\mathbf{R}_{aa} + \lambda\,\mathbb{I}_M)^{-1}\mathbf{A}^\top) \tag{16}$$

.

3 Results

The results are obtained for the auxetic form data for 8 datasets using both types of regularisation. Ridge traces are shown for values of λ from 10^{-6} to 10^{0}. While the RLS is usually used for the normalised (scaled and centred) data, results for the raw

4 Future work

- Regularised LS for the direct estimation of B from joint datasets. Long time-series and polynomial structure of the regressors may also lead to collinearity problems. It's not clear yet whether the data needs to be normalised separately.
- Confidence and covariance regions to demonstrate bias-variance trade off.
- Multiple methods for LASSO regression can be investigated.
- Cross-validation and other methods to select the regularisation term.