Sparse parameter estimation of LTI models with $l^p$ sparsity using genetic algorithm

Vikram Saini* and Lillie Dewan

Electrical Engineering Department,
National Institute of Technology,
Kurukshetra-136119, Haryana, India
Email: ivikramsaini@gmail.com
Email: l_dewan@nitkkr.ac.in
*Corresponding author

Abstract: Sparse optimisation for the identification of parametric linear model structure is equivalent to the estimation of the parameter vector. After relaxing the assumption on the order of the system, sparse optimisation techniques can be utilised to find the optimal model. This paper proposes an optimisation method to find the sparse parameter estimates. For this purpose, $l^p$ norm ($0 < p < 1$) penalty of parameter vector is added to the quadratic loss function which is further minimised using genetic algorithm. For the model structures other than ARX, a simulation model is realised using conditions on the quadratic simulation error. A real coded genetic algorithm is used to minimise the simulation error model. Simulation results are given for the ARX and output error model structures to show the effectiveness of the simulation error model method.

Keywords: sparse optimisation; genetic algorithm; $l^p$ sparsity measure; simulation error model.


Biographical notes: Vikram Saini is currently pursuing his PhD in the Department of Electrical Engineering, National Institute of Technology, Kurukshetra, Haryana, India. His area of research includes identification of linear and non-linear systems, and control systems theory.

Lillie Dewan is a Professor with the Electrical Engineering Department, National Institute of Technology, Kurukshetra, Haryana, India. Her area of interest includes identification, control theory (adaptive, optimal, robotics), signal processing and communication.

1 Introduction

System identification deals with the parameter estimation of mathematical models from observed input-output data (Laamiri et al., 2015; Ahmad, 2015; Thai et al., 2016). In identification literature, the popular approach is the ordinary least squares (OLS) or the minimisation of the residual squared error (Soderstrom and Stoica, 1989; Ljung, 1999; Schimmack and Mercorelli, 2016). However, the prediction accuracy of estimates has low bias but large variance. This can be improved by selecting less number of parameters, which correspondingly increases the structural bias. So, the estimation problem of parametric models is to optimise the bias-variance trade-off. The bias-variance trade-off is equivalent to the transformation of high dimensional model to low dimensional model in the optimum sense or to find the sparse solutions. The sparse estimation problem has been earlier studied by Jeffs and Gunsay (1993) and Harikumar and Bresler (1996) to find the estimates of signals and systems. The motivation for such an approach is that a lower order model is sometimes sufficient to represent the signals and systems efficiently.

Many methods have been suggested for the model order selection such as subset selection method in which a best subset is selected within the subsets, validated on some cross-validation technique (Akaike, 1974; Schwarz, 1978). However, the method is computationally intensive if the parameters are very large in number because these selections are not regularised. So, the sparse solutions in the form of regression with regularised techniques were proposed to address the problems of un-regularised methods. The idea is to minimise the squared error subject to some constraints on the parameter vector. The regularised methods provide a computationally feasible way in the complexity of one regression. Different regularised methods have been suggested, e.g., ridge regression (Tikhonov and Arsenin, 1977), non-negative Garrotte (NNG) (Breiman, 1995), least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) and elastic net method (ENM) (Zou and Hastie, 2005). However, Fu (1998) proposed a non-convex measure $l^p$ ($0 < p < 1$) to generate the sparsest solutions.
In case of sparse optimisation, some transfer function models give rise to non-convex functions such as output-error, ARMAX and Box-Jenkins models. In addition, the noise sequence is generally unavailable prior to the estimation. So, it is very difficult to realise the optimum model structure. This gives us the opportunity to use evolutionary approach to explore sparse estimation problem which under certain conditions can effectively estimate the sparse parameter vector. The stochastic search algorithms (especially genetic algorithm) have been used earlier for the statistical modelling (Minerva and Paterlini, 2002; Bao and Cassandras, 1996) and chaotic signals estimation (Han and Chang, 2011). The genetic algorithm is a search algorithm based on the conjecture of natural selection and genetics. Hence, there is no need for computation of derivatives or other auxiliary functions to guide its search. This paper proposes the genetic algorithm for the sparse estimation of linear time invariant (LTI) systems using the simulation error (SE) model with the $\ell^p$ ($0 < p < 1$) penalty.

After the brief introduction, problem is described in Section 2. In Section 3, the Lagrangian formulation for sparse estimation is given. Section 4 establishes the SE model with certain conditions. In Section 5, two numerical examples are presented to evaluate the effectiveness of the proposed method. Finally, conclusions are drawn in Section 6.

## 2 Problem description

Let $u(k), y(k)$ and $v(k) \in \mathbb{R}$ be the input, output and noise of the physical system respectively. A stable discrete LTI dynamic system is described as (Ljung, 1999)

$$y(k) = \frac{B(q^{-1})}{A(q^{-1})}u(k) + \frac{D(q^{-1})}{C(q^{-1})}v(k)$$

(1)

where

$$A(q^{-1}) = 1 + \sum_{i=1}^{\infty} a_i q^{-i}, B(q^{-1}) = \sum_{i=1}^{\infty} b_i q^{-i}$$

$$C(q^{-1}) = 1 + \sum_{i=1}^{\infty} c_i q^{-i}, D(q^{-1}) = 1 + \sum_{i=1}^{\infty} d_i q^{-i}$$

(2)

are polynomials in the backward shift operator $q^{-1}$. The generalised discrete model structure defined by:

$$y(k) = \frac{\hat{B}(q^{-1})}{\hat{A}(q^{-1})}u(k) + \frac{\hat{D}(q^{-1})}{\hat{C}(q^{-1})}v(k)$$

(3)

can be utilised in the polynomial form to represent (1).

Where as usual,

$$\hat{A}(q^{-1}) = 1 + \sum_{i=1}^{\infty} \hat{a}_i q^{-i}, \hat{B}(q^{-1}) = \sum_{i=1}^{\infty} \hat{b}_i q^{-i}$$

$$\hat{C}(q^{-1}) = 1 + \sum_{i=1}^{\infty} \hat{c}_i q^{-i}, \hat{D}(q^{-1}) = 1 + \sum_{i=1}^{\infty} \hat{d}_i q^{-i}$$

(4)

By setting one or more polynomials in (3) to unity can efficiently represent different model structures such as ARX, ARMAX, output-error and Box-Jenkins (Soderstrom and Stoica, 1989). The prediction output for the approximate model structure (OE) of (1) is

$$\hat{y}(k|k-1) = w(k, \theta) = -\sum_{i=1}^{\infty} \hat{a}_i w(k-i) + \sum_{i=1}^{\infty} \hat{b}_i u(k-i)$$

(5)

where $\hat{y}(k|k-1)$ is the one step ahead prediction of $y$ at $k$th instant using previous data of $y$ up to $(k-1)$ samples. Define $a \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, $\phi_i(k) \in \mathbb{R}^n$ and $\phi(k) \in \mathbb{R}^n$ as

$$a = [\hat{a}_1 \hat{b}_1 \ldots \hat{a}_n \hat{b}_n], b = [\hat{b}_1 \hat{b}_2 \ldots \hat{b}_n], \phi_i(k) = [u(k-1) - w(k-2) \ldots - w(k-n)],$$

(6)

$$\phi(k) = [u(k-1) \ldots u(k-n)]$$

Also define the stacked information matrices $\phi \in \mathbb{R}^{L \times n}$ and $\Phi \in \mathbb{R}^{L \times n}$ as

$$\phi := [\phi_1(1) \phi_2(2) \ldots \phi_L(L)], \Phi := [\phi_1(1) \phi_2(2) \ldots \phi_L(L)]$$

(7)

Then, the model (5) can be expressed in linear regression form as

$$Y = \Phi^T a + \Phi^T b$$

and the parameters estimates can be obtained by the following minimisation

$$\arg \min_{\theta} \text{Loss}(\theta) := \arg \min_{\theta} \frac{1}{L} \left\| Y - \Phi^T \theta \right\|^2$$

(8)

where $\theta = [a^T \ b^T] \in \mathbb{R}^{n+1}$ and $\psi = [\phi_1^T \ \phi_2^T \ \ldots \ \phi_L^T] \in \mathbb{R}^{(n+1) \times L}$. Let $\theta' \in \mathbb{R}^{n}$ represents the parameter vector of system and $\hat{\theta} \in \mathbb{R}^n$ is the estimate corresponding to (8). Then either $\hat{\theta}$ results into a false minima or inconsistent if $\text{dim}(\theta') < \text{dim}(\theta')$ (model order is unknown). Also, if the number of elements in $\text{Supp}(\theta') < \text{dim}(\theta')$ ($\text{Supp}(\theta') = \{i \in \mathbb{Z}^+ : \theta'_i \neq 0\}$) then $\theta'$ is a sparse vector. But, OLS solution (8) never estimates zero parameter values (Tikhonov and Arsenin, 1977; Seidman, 1980).

For $k \leq 0$, it is assumed that $u(k), y(k)$ and $v(k)$ are zero. The goal is to find the SE model with a minimal prior knowledge, that best fits the measured dataset $(u(k), y(k))$ for $k \in T, \text{Card}(T) = L$ where $T \subset \mathbb{Z}^+$, which is known as ill posed problem (Tikhonov and Arsenin, 1977).

Note: In this paper, model order is unknown and it is assumed that model structure is known prior to estimation process. However, as a relaxation, it will be shown that
3 Penalised loss function with sparsity measure

In this paper, non-convex measure $p$ ($0 < p < 1$) is analysed (Fu, 1998; Frank and Friedman, 1993). Let $\mu = (\mu_1, \mu_2, \ldots, \mu_n) \in \mathbb{R}^n$, $\mu_i > 0$ for $i = 1, 2, \ldots, n$ and $\Omega$ be the set of vectors $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$ in $\mathbb{R}^n$ such that $|\theta|^p < \mu$, for $i = 1, 2, \ldots, n$ where $0 < p < 1$.

Then for non-convex measure $p$, minimisation problem is stated as

$$\min_{\theta \in \Omega} \text{Loss}(\theta)$$

subject to $f(\theta) < \mu$. Consider the Lagrangian function

$$L(\theta, \lambda) = \text{Loss}(\theta) + \lambda^T(f(\theta) - \mu)$$

then, the dual Lagrangian problem is given by

$$\max_{\lambda \in \mathbb{R}^n} g(\lambda)$$

subject to $\lambda_i > 0$, for $i = 1, 2, \ldots, n$

where $g(\lambda) = \inf_{\theta \in \mathbb{R}^n} L(\theta, \lambda)$ is the dual function. Now it can be shown that in case of large measurement samples of data available and under weak duality theorem, (9) is equivalent to (11) and $g(\lambda) \leq \text{Loss}^* (\text{loss}^*$ is the optimum value). The duality gap between primal and dual function is $\text{Loss}^*(\theta) - g(\lambda)$ ($g^*$ is the dual optimum values). If $\text{Loss}^*(\theta) - g(\lambda) \leq \varepsilon$ then as a consequence $\text{Loss}(\theta) - \text{Loss}^*(\theta) \leq \varepsilon$ is feasible.

The problem (11) is not computationally attractive as this requires finding $n$ variables $\lambda_i$’s. To overcome this, the problem can be redefined with a single constrained $\lambda' \in \mathbb{R}$ as

$$\arg \min_{\lambda' \in \mathbb{R}} \frac{1}{L} \left[ \|Y - y^T \theta\|^2_2 + \lambda' \|\theta\|_p \right] \text{ for } 0 < p < 1$$

For the choice of $\lambda'$, theoretically $\lambda' = O(\sqrt{L})$ is necessary, whereas $\lambda' = O(L^{p/2})$ is sufficient (Knight and Fu, 2000). The vector space $\mathbb{R}^n$ with norm $\|\cdot\|_p$ is neither convex nor concave and possesses many strong local minima because as $p \to 0$, function $\|\cdot\|_p$ is more spiked. The objective function (12) is not differentiable when $\theta$ is a sparse vector, which causes the failure of standard gradient methods. Frank and Friedman (1993) and Huang et al. (2008) have provided approximate solutions to remove the problem of singularity, which is useful to find one step estimates. But, by using evolutionary algorithm, $\|\cdot\|_p$ can be considered as a sparsity measure, which do not require approximation for the sparse solutions of LTI systems.

Note: $\|\cdot\|_p$ with $0 < p < 1$ is actually not a norm as the triangle inequality doesn’t hold and it is consistent with the definition of quasi norm, though distance function $d(\cdot, \cdot) = \|\cdot - \cdot\|_p^2$ is a well defined metric (Pietsch, 1981).

4 Estimation as an optimisation problem

As seen in Section 3, identification of LTI dynamical models along with sparsity measure is highly non-convex in nature and therefore is a hard combinatorial problem. The difficulty with non-convex function is that a stationary point does not necessarily be a global minimum. Therefore to get a sparse solution, some evolutionary algorithm can be used.

4.1 SE model

The SE approach has been used for the estimation of models (Bao and Cassandras, 1996; Connolly et al., 2007). Assume that the dataset $D_L = \{(u_k, y(k)), k \in T, \text{Card}(T) = L$ and $T \subset \mathbb{Z}^+$} is available from the system (1). Let $\hat{y}(k)$ is the simulated output of the model

$$M(\hat{\theta}) : \hat{y}(k) = f_1(q^{-1}, \hat{y}(k), a) + f_2(q^{-1}, u(k), b)$$

where $f_1$ and $f_2$ are linear scalar functional, and assuming $f_1(0, \hat{y}(k), a) = 0$ and $f_2(0, u(k), b) = 0$. Further, the SE defined by $\varepsilon_{SE} = y(k) - \hat{y}(k)$ may in general be unbounded in the negative region for asymptotically unstable models. A remedy to this is to use some positive function for the SE, e.g., quadratic function. Note that the model (13) will generate the same set of simulated outputs for the ARX, ARMAX and OE model structure. Then the minimisation of SE may trap in the bias-variance trade-off problem resulting in non-sharp sparse estimates if the true system belongs to OE or ARMAX system. Define $e(k)$ as

$$e(k) = \frac{\hat{D}(q^{-1})}{C(q^{-1})} \gamma(k)$$

Further, make the following assumptions on system (1).

$$A_1 |\gamma(k)| < \gamma_i \text{ for some } \gamma_i \in \mathbb{R}^+, \text{ and for all } k.$$  

$$A_2 |e(k)| < \gamma_2 \text{ for some } \gamma_2 \in \mathbb{R}^+, \text{ and for all } k.$$  

In the OE case, the solution can be obtained iteratively by using a filter (Steiglitz and Mcbridge, 1965) such that the SE can be expressed as

$$\varepsilon_{SE} = \left( \frac{1}{A(q^{-1}, a)} \right) (y(k) - \hat{y}(k|\theta))$$

Alternatively, under certain conditions the SE model can provide parameters estimates. For this, define $Y = [y(1) y(2) \ldots y(L)]^T \in \mathbb{R}^{1 \times L}$, $\hat{Y} = [\hat{y}(1|\theta) \hat{y}(2|\theta) \ldots \hat{y}(L|\theta)]^T \in \mathbb{R}^{1 \times L}$, and $E = [e(1) e(2) \ldots e(L)]^T$. Denote the vectors $Y, \hat{Y}$ and $E$ by $x, \hat{x}$ and $e$, respectively. Then using triangle inequality for normed spaces we get

$$\|x - \hat{x}\|_2 \leq \|x - e\|_2 \leq \|x - \hat{x}\|_2 + \|e\|_2$$
Lemma 4.1.1: If $\| x - \hat{x} \|_2 > \frac{\| e \|_2}{2}$, then $\exists g(x - \hat{x}) : \Omega \rightarrow \mathbb{R}^*$ such that

$$\| x - \hat{x} \|_2^2 - g(x - \hat{x}) \leq \| x - \hat{x} - e \|_2^2$$

and the equality holds if $\| x - \hat{x} \|_2 = \| e \|_2^2$.

Proof: Squaring both sides of the LHS inequality of (16) and simplifying we get

$$0 \leq \| x - \hat{x} \|_2^2 + \| e \|_2^2 - 2 \| x - \hat{x} \|_2 \| e \|_2 \leq \| x - \hat{x} - e \|_2^2$$

(17)

Now define $f(x - \hat{x}) = -\| e \|_2^2 + 2 \| x - \hat{x} \|_2 \| e \|_2$ where $\| x - \hat{x} \|_2 > \| e \|_2$, then using (17) we get

$$\| x - \hat{x} \|_2^2 - g(x - \hat{x}) \leq \| x - \hat{x} - e \|_2^2$$

Now for the desired solution, the condition $\| x - \hat{x} \|_2 > \frac{\| e \|_2^2}{2}$ could be weakened to $\| x - \hat{x} \|_2 \geq \| e \|_2$. Note that $g$ is a function of $x - \hat{x}$ and since norm is a continuous function therefore, with the decrease of $\| x - \hat{x} \|_2$, the function $\| x - \hat{x} - e \|_2$ and $g(\cdot)$ will also decrease. If $\| x - \hat{x} \|_2 = \| e \|_2$, then equality holds in (17) and $(x - \hat{x}) = \| e \|_2^2$.

This completes the proof: ■

Under the influence of Lemma 4.1.1, a suitable minimisation function is given as

$$\| x - \hat{x} \|_2^2 - g(x - \hat{x}) \text{ for } \| x - \hat{x} \|_2 \geq \| e \|_2$$

(18)

This function is suitable as its derivative equating to zero will give the unbiased least square estimates. Also, it is well defined in combination with $\ell^p$ with $0 < p < 1$. Because $\lambda \| \theta \|_p$ can be added on both sides of the inequality (18). But as $\| e \|_2$ is unknown this function is almost un-realisable. However, this could be approximated to some other realisable function under certain conditions.

Lemma 4.1.2: If $0 < \beta < 1$. Then $\exists \beta$, $0 < \beta < 1$ such that

$$1 - \beta = -b^2 + 2b.$$

Proof: Let $\alpha = -b^2 + 2b - b^2 + 2b - 1 = -(b - 1)^2 + 1 = -(1 - b)^2 + 1$.

Since $0 < b < 1$, therefore $0 < 1 - b < 1$. Hence, $0 < (1 - b)^2 + 1$ and $-1 < -(1 - b)^2 < 0$. From this it follows that $0 < 1 - (1 - b)^2 < 1$, i.e., $0 < \alpha < 1$.

Let $\alpha = 1 - \beta$, then $1 - \beta = -b^2 + 2b$, where $0 < \beta < 1$.

This completes the proof: ■

Now using the definition of $g(\cdot)$

$$g(x - \hat{x}) = \frac{-\| e \|_2^2 + 2\| x - \hat{x} \|_2 \| e \|_2}{\| x - \hat{x} \|_2^2}$$

(19)

Under the condition $\| x - \hat{x} \|_2 > \| e \|_2$ (however, a lower bound is $\| x - \hat{x} \|_2 > \| e \|_2/2$), if $b = \| e \|_2/\| x - \hat{x} \|_2$, then Lemma 4.1.2 shows that

$$(1 - \alpha_\beta)(x - \hat{x})^2 = \| x - \hat{x} \|_2^2 - g(x - \hat{x})$$

(20)

where $\beta_\alpha = (1 - a\alpha)$. The plot of $\beta_\alpha$ as a function of $\| e \|_2$ and $\| x - \hat{x} \|_2$ is shown in Figure 1, which clearly shows that the value of $\beta_\alpha \rightarrow 0$ as $\| x - \hat{x} \|_2 \rightarrow \| e \|_2$. The plot of $\beta_\alpha$ and $\alpha_\beta$ as a function of $b = \| e \|_2/\| x - \hat{x} \|_2$ is shown in Figure 2(a) and Figure 2(b), respectively.

Figure 1 Plot of beta ($\beta$) as a function of $\| e \|_2$ and $\| x - \hat{x} \|_2$ (see online version for colours)

From Figure 2, it is observed that

$$\lim_{\| x - \hat{x} \|_2 \rightarrow \| e \|_2} \frac{\| e \|_2}{\| x - \hat{x} \|_2} \rightarrow 1,$$

when $\beta_\alpha \rightarrow 0$ and $\alpha_\beta \rightarrow 1$. Note that when $\alpha_\beta \rightarrow 1$, the quadratic function vanishes, leading to a zero solution. So, an exponential function with negative argument of $\| x - \hat{x} \|_2$ for $\alpha_\beta$ might be considered for the optimisation. Alternatively, a constant value of $\alpha$ strictly less than one will give consistent estimates. In addition, for the combinatorial optimisation with $\ell^p$ penalty, the value of $\lambda^p$ must be accordingly changed. Then the sparse solution is obtained by the following minimisation.
arg min_{\hat{\theta} \in \mathbb{R}^n} \left( (1 - \alpha) \| y - \hat{y} \|^2_2 + (1 - \alpha) \lambda^2 \| \hat{\theta} \|_p \right) \quad (21)

for 0 < p < 1

Note: In case of OE assumption $A_2$ requires $A_1$ and the polynomial $A(q^{-1}, a)$ must be a Hurwitz polynomial, which is an inherent standard assumption in identification of linear time-invariant systems.

Figure 2  Plot of beta ($\beta$) as a function of $b = \frac{\| e \|_2}{\| x - \hat{x} \|_2}$

(b) Plot of alpha ($\alpha$) as a function of $b = \frac{\| e \|_2}{\| x - \hat{x} \|_2}$

(see online version for colours)

4.2 Genetic operator minimisation

Genetic algorithm has global and parallel search ability and is appropriate for solving complex hard combinatorial problems (Minerva and Paterlini, 2002; Goldberg, 2006; Yang and Douglas, 1998). Let there be a function $J(\hat{\theta}) : \Omega \rightarrow \mathbb{R}$ which is to be minimised over a domain $\Omega$.

Step 1 Initialisation – Initialise the population size and selects $m$ elements $\hat{\theta}_i$, $i = 1, \ldots, m$ from $\Omega$ that randomly represents the individuals for the first generation.

Step 2 Scaling and selection: Compute the fitness function for the $m$ elements $\hat{\theta}_i$ as $F(\hat{\theta}_i) = J(\hat{\theta}_i)$ and the survival probability $p_\alpha$. Select the fit individuals from the population using the stochastic selection method, such that $Z = \{ \hat{\theta}_i = \hat{\theta} : \hat{\theta}_i$ is most fit, $Z \subset \Omega \}$.

Step 3 Reproduce: Compute population for the next generation using single point crossover operator and Gaussian mutation operator (Belfiore and Esposito, 1998).

Step 4 Repeat the Steps 2 and 3 until some criterion is fulfilled.

It is reasonable to expect that when the algorithm terminates, the final solution is usually converge to the exact solution. But as a case of numerical imp-recision it is usually close to the exact solution. A heuristic could be used to find the support of parameter vector that represent the non-zero elements. The idea is to assume that zero valued components are at some distance from the exact solution (exact solution is zero) (Bredies et al., 2014).

$\text{Supp}(\hat{\theta}) = \{ i : \| \hat{\theta}_i \|_\infty > 10^{-2}, i = 1, 2, \ldots, \text{dim}(\hat{\theta}) \} \quad (22)$

5 Computer simulation

In this section, simulation results are presented to evaluate the performance of the proposed approach to find the sparse estimates. For the sake of brevity, two examples are presented with different model structures. In Example 1 model structure is ARX, whereas in Example 2 it is OE structure. For both the examples, a dataset with $L = 600$ is gathered from the system with $u$ being drawn from zero mean and unit variance with normal distribution $N(0, 1)$. The signal $v(k)$ is a white noise with Gaussian distribution $N(0, \sigma_v^2)$ with $\sigma_v < 1$. The numerical value of $\lambda$ is chosen according to the cross validation using the best fit rate (BFR) criterion. The numerical value of $p$ for $p$ penalty is taken as 0.1.

Example 1: Consider the following discrete ARX system given as:

$y(k) = b_1 u(k - 1) + b_2 u(k - 2) - a_1 y(k - 1)$

where $a_1 = 0.55$, $a_2 = 0.8$, $b_1 = 0.15$, $b_2 = -0.35$. For the sparse estimation of given ARX system, the SE model with orders $n_a = n_b = 4$ is considered. So in the given model, the sparse solution contains four non-zero and four zero parameters. To show the effectiveness of SE model with a contraction operator, different cases of quadratic function with and without the constraint $\| \hat{\theta} \|_p$, for different values of $\alpha$ are considered. In all cases, the algorithm is initialised
Sparse parameter estimation of LTI models with \( p \) sparsity using genetic algorithm

without bounds on the parameters, keeping the population size equal to 60. For the termination of the algorithm, the number of generations is taken 150. The best model is chosen according to the BFR criterion and, \( \lambda' = 32 \) is obtained. The obtained parameters estimates for each case are shown in Table 1 and graphically compared in Figure 3.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>( a_1 )</th>
<th>( a_2 )</th>
<th>( a_3 )</th>
<th>( a_4 )</th>
<th>( b_1 )</th>
<th>( b_2 )</th>
<th>( b_3 )</th>
<th>( b_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quadratic w/o ( p )</td>
<td>0.4017</td>
<td>0.7481</td>
<td>–0.0993</td>
<td>0.0353</td>
<td>0.1536</td>
<td>–0.3462</td>
<td>0.0736</td>
<td>0.0151</td>
</tr>
<tr>
<td>Quadratic w/o ( p \alpha = 0.15 )</td>
<td>0.8397</td>
<td>0.8533</td>
<td>0.1917</td>
<td>–0.0784</td>
<td>0.1958</td>
<td>–0.2247</td>
<td>–0.0276</td>
<td>0.0965</td>
</tr>
<tr>
<td>Quadratic with ( p )</td>
<td>0.4700</td>
<td>0.7561</td>
<td>–0.0709</td>
<td>0.0050</td>
<td>0.1104</td>
<td>–0.3847</td>
<td>–0.0022</td>
<td>0.0050</td>
</tr>
<tr>
<td>Quadratic with ( p \alpha = 0.15 )</td>
<td>0.6546</td>
<td>0.8554</td>
<td>0.0804</td>
<td>–0.0035</td>
<td>0.1806</td>
<td>–0.3152</td>
<td>–0.0067</td>
<td>–0.0030</td>
</tr>
<tr>
<td>Quadratic with ( p \alpha = 0.50 )</td>
<td>0.7064</td>
<td>0.8722</td>
<td>0.0981</td>
<td>–0.0044</td>
<td>0.1741</td>
<td>–0.3042</td>
<td>0.0016</td>
<td>–0.0014</td>
</tr>
<tr>
<td>True value</td>
<td>0.55</td>
<td>0.80</td>
<td>0</td>
<td>0</td>
<td>0.15</td>
<td>–0.35</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Table 2** Estimates of parameters with different cases of OE model structure

<table>
<thead>
<tr>
<th>Parameters</th>
<th>( a_1 )</th>
<th>( a_2 )</th>
<th>( a_3 )</th>
<th>( a_4 )</th>
<th>( b_1 )</th>
<th>( b_2 )</th>
<th>( b_3 )</th>
<th>( b_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quadratic w/o ( p )</td>
<td>0.5727</td>
<td>0.8080</td>
<td>0.1456</td>
<td>0.0342</td>
<td>0.2300</td>
<td>–0.1625</td>
<td>0.1823</td>
<td>0.2278</td>
</tr>
<tr>
<td>Quadratic w/o ( p \alpha = 0.15 )</td>
<td>0.6213</td>
<td>0.8966</td>
<td>0.1156</td>
<td>0.0483</td>
<td>0.1992</td>
<td>–0.2215</td>
<td>0.1174</td>
<td>0.0780</td>
</tr>
<tr>
<td>Quadratic with ( p ) prefILTER</td>
<td>0.4799</td>
<td>0.7480</td>
<td>–0.0820</td>
<td>–0.0088</td>
<td>0.1995</td>
<td>–0.2856</td>
<td>0.0914</td>
<td>0.0563</td>
</tr>
<tr>
<td>Quadratic with ( p ) prefILTER and ( \alpha = 0.15 )</td>
<td>0.4752</td>
<td>0.7441</td>
<td>–0.0272</td>
<td>0.0011</td>
<td>0.1029</td>
<td>–0.3848</td>
<td>0.0012</td>
<td>0.0044</td>
</tr>
<tr>
<td>Quadratic with ( p \alpha = 0.15 )</td>
<td>0.6569</td>
<td>0.8374</td>
<td>0.0931</td>
<td>–0.0057</td>
<td>0.2022</td>
<td>–0.2667</td>
<td>0.0498</td>
<td>0.0840</td>
</tr>
<tr>
<td>Quadratic with ( p \alpha = 0.50 )</td>
<td>0.7142</td>
<td>0.9322</td>
<td>0.0914</td>
<td>0.0026</td>
<td>0.1974</td>
<td>–0.2715</td>
<td>–0.0067</td>
<td>–0.0441</td>
</tr>
<tr>
<td>Quadratic with ( p \alpha = 0.80 )</td>
<td>0.6885</td>
<td>0.8913</td>
<td>0.0513</td>
<td>–0.0005</td>
<td>0.1734</td>
<td>–0.2967</td>
<td>0.0010</td>
<td>–0.0042</td>
</tr>
<tr>
<td>Quadratic with ( p \alpha = 0.80 )</td>
<td>0.5293</td>
<td>0.7918</td>
<td>–0.0069</td>
<td>0.0001</td>
<td>0.1777</td>
<td>–0.3179</td>
<td>–0.0024</td>
<td>0.0088</td>
</tr>
<tr>
<td>True value</td>
<td>0.55</td>
<td>0.80</td>
<td>0</td>
<td>0</td>
<td>0.15</td>
<td>–0.35</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Figure 4  Parameter estimates with different cases of OE model structure (see online version for colours)

Table 3  Estimate of parameters with best and worst case out of 10 runs

<table>
<thead>
<tr>
<th>Parameters</th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$a_3$</th>
<th>$a_4$</th>
<th>$b_1$</th>
<th>$b_2$</th>
<th>$b_3$</th>
<th>$b_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARX (variable $\alpha$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best</td>
<td>0.5509</td>
<td>0.8063</td>
<td>-0.0006</td>
<td>0.0001</td>
<td>0.1846</td>
<td>-0.3403</td>
<td>0.0087</td>
<td>0.0028</td>
</tr>
<tr>
<td>Worst</td>
<td>0.5427</td>
<td>0.8002</td>
<td>-0.0095</td>
<td>-0.0081</td>
<td>0.1720</td>
<td>-0.2858</td>
<td>-0.0023</td>
<td>0.0071</td>
</tr>
<tr>
<td>OE (variable $\alpha$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best</td>
<td>0.5284</td>
<td>0.7834</td>
<td>0.0023</td>
<td>-0.0091</td>
<td>0.1294</td>
<td>-0.3861</td>
<td>-0.0030</td>
<td>-0.0089</td>
</tr>
<tr>
<td>Worst</td>
<td>0.4920</td>
<td>0.8701</td>
<td>0.0086</td>
<td>0.0032</td>
<td>0.0989</td>
<td>-0.3608</td>
<td>-0.0031</td>
<td>-0.0010</td>
</tr>
<tr>
<td>True value</td>
<td>0.55</td>
<td>0.80</td>
<td>0</td>
<td>0</td>
<td>0.15</td>
<td>-0.35</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

First case is considered without the constraint $||\theta||_p$, keeping value of $\alpha = 0$. However, in this case, none of the sparse parameters is identified. But by increasing the value of $\alpha$ to 0.15, the estimates tend towards sparsity. Second case is considered with the iteratively pre-filter technique using SE (15). For the pre-filter case, it is not necessary to introduce $\alpha$, however, it provides a consistent support of the sparse parameters. Results with a single value of $\alpha = 0.15$, are presented in this paper, whereas it is observed that as the value of $\alpha \to 1$, it gives a better true support. A third case is considered with the constraint $||\theta||_p$ defined in (20) and using three values of $\alpha$ (0.15, 0.50 and 0.80). It is seen that when $\alpha \to 1$, minimisation (20) leads to more sparse solutions for the given constraints.

Example 2: Consider the following discrete OE system given as:

$$y(k) = \frac{\hat{\theta}(q^{-1}, \theta)}{\hat{A}(q^{-1}, \alpha)} u(k) + v(k)$$

where $a_1 = 0.55$, $a_2 = 0.80$, $b_1 = 0.15$, $b_2 = -0.35$. The SE model with orders $n_u = n_y = 4$ is considered for the sparse estimation. The genetic algorithm structure is the same as taken in the first example. The best model is chosen according to the BFR criterion and, $\lambda' = 40$ is obtained. Different cases have been considered. The obtained parameters estimates for each case is shown in Table 2 and graphically compared in Figure 4.

From Table 1, it can be seen that the minimisation with and without the constraint $||\theta||_p$, the value of redundant parameters is effectively lower than the threshold and it represents the parameters that must be zero in the model. For the case of minimisation of SE with sparsity measure, increased value of $\alpha$ yields a better estimation of the true parameters support. In each case, some redundant parameters having higher value than the threshold are also estimated by the algorithm. However, a two stage iterative procedure tends to produce more accurate parameter support. After identifying the parameter support, the non-zero parameters can be re-estimated to find the model that best represents the system in the optimal sense.
6 Conclusions

This paper presents sparse optimisation of LTI model structure using the Lagrangian function. For the sparse parameter estimation, $\ell^p$ norm ($0 < p < 1$) of parameters vector is added to quadratic loss function, which is further minimised by genetic algorithm. Owing to non-convexity and singularity at the origin of $\ell^p$ norm, it is almost unrealisable when standard gradient methods are used. In addition, owing to the presence of noise in the modelling, model structure OE gives rise to a non-convex quadratic error function. Therefore, a SE model is iteratively optimised by providing conditions on the quadratic error for the sparse estimation. Simulation results of ARX and OE model structures using the SE model show the effectiveness of proposed scheme.

References


