Segmentation of ARX-models Using Sum-of-Norms Regularization

Henrik Ohlsson, Lennart Ljung, Stephen Boyd
Division of Automatic Control
E-mail: ohlsson@isy.liu.se, ljung@isy.liu.se, boyd@stanford.edu

25th April 2010

Report no.: LiTH-ISY-R-2941
Accepted for publication in Automatica

Address:
Department of Electrical Engineering
Linköpings universitet
SE-581 83 Linköping, Sweden

WWW: http://www.control.isy.liu.se
Abstract
Segmentation of time-varying systems and signals into models whose parameters are piecewise constant in time is an important and well studied problem. It is here formulated as a least-squares problem with sum-of-norms regularization over the state parameter jumps, a generalization of $\ell_1$-regularization. A nice property of the suggested formulation is that it only has one tuning parameter, the regularization constant which is used to trade off fit and the number of segments.

Keywords: segmentation, regularization, ARX-models
Segmentation of ARX-models Using Sum-of-Norms Regularization

Henrik Ohlsson a, Lennart Ljung a, Stephen Boyd b

a Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden
b Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA

Abstract

Segmentation of time-varying systems and signals into models whose parameters are piecewise constant in time is an important and well studied problem. It is here formulated as a least-squares problem with sum-of-norms regularization over the state parameter jumps, a generalization of \( l_1 \)-regularization. A nice property of the suggested formulation is that it only has one tuning parameter, the regularization constant which is used to trade off fit and the number of segments.

Key words: segmentation, regularization, ARX-models

1 Model Segmentation

Estimating linear regression models

\[
y(t) = \varphi^T(t)\theta
\]  

(1)

is probably the most common task in system identification. It is well known how ARX-models

\[
y(t)+a_1y(t-1)+\cdots+a_ny(t-n) = b_1u(t-nk-1)+\cdots+b_mu(t-nk-m)
\]

(2)

with inputs \( u \) and outputs \( y \) can be cast in the form (1). Time-series AR models, without an input \( u \) are equally common.

The typical estimation method is least-squares,

\[
\hat{\theta}(N) = \arg \min_{\theta} \sum_{t=1}^{N} \|y(t) - \varphi^T(t)\theta\|^2
\]

(3)

where \( \cdot \) denotes the Euclidean or \( l_2 \) norm.

1.1 Common Case

A common case is that the system (model) is time-varying:

\[
y(t) = \varphi^T(t)\theta(t)
\]

(4)

A time-varying parameter estimate \( \hat{\theta} \) can be provided by various tracking (on-line, recursive, adaptive) algorithms. A special situation is when the system parameters are piecewise constant, and change only at certain time instants \( t_k \) that are more or less rare:

\[
\theta(t) = \theta_k, \quad t_k < t \leq t_{k+1}
\]

(5)

This is known as model or signal segmentation and is common in e.g. signal analysis (like speech and seismic data), failure detection and diagnosis. There is of course a considerable literature around all this and its ramifications, e.g. [18], [15], [3].

The segmentation problem is often addressed using multiple detection techniques, multiple models and/or Markov models with switching regression, see, e.g. [17], [27], [6]. The function segment for the segmentation problem in the System Identification Toolbox [19], is based on a multiple model technique [2].

2 Our Method

We shall in this contribution study the segmentation problem from a slightly different perspective. If we al-
low all the parameter values in (4) to be free in a least-squares criterion we would get
\[\min_{\theta(t), t=1,\ldots,N} \sum_{t=1}^{N} \|y(t) - \varphi^T(t)\theta(t)\|^2.\]

Since the number of parameters then exceeds or equals the number of observations we would get a perfect fit, at the price of models that adjust in every time step, following any momentary noise influence. Such a grossly over-fit model would have no generalization ability, and so would not be very useful.

2.1 Sum-of-Norms Regularization

To penalize model parameter changes over time, we add a penalty or regularization term (see e.g. [8, p.308]) that is a sum of norms of the parameter changes:
\[\min_{\theta(t), t=1,\ldots,N} \sum_{t=1}^{N} \|y(t) - \varphi^T(t)\theta(t)\|^2 + \lambda \sum_{t=2}^{N} \|\theta(t) - \theta(t-1)\|_{\text{reg}},\]

where \(\|\cdot\|_{\text{reg}}\) is the norm used for regularization, and \(\lambda\) is a positive constant that is used to control the trade-off between model fit (the first term) and time variation of the model parameters (the second term). The regularization norm \(\|\cdot\|_{\text{reg}}\) could be any vector norm, like \(\ell_1\) or \(\ell_p\), but it is crucial that it is a sum of norms, and not a sum of squared norms, which is the more usual Tychonov regularization.

When the regularization norm is taken to be the \(\ell_1\) norm, i.e., \(\|z\|_1 = \sum_{k=1}^{n} |z_k|\), the regularization in (6) is standard \(\ell_1\) regularization of the least-squares criterion. Such regularization has been very popular recently, e.g. in the much used Lasso method, [26] or compressed sensing [11,9]. There are two key reasons why the parameter fitting problem (6) is attractive:

- It is a convex optimization problem, so the global solution can be computed efficiently. In fact, its special structure allows it to be solved in \(O(N)\) operations, so it is quite practical to solve it for a range of values of \(\lambda\), even for large values of \(N\).
- The sum-of-norms form of the regularization favors solutions where “many” (depending on \(\lambda\)) of the regularized variables come out as exactly zero in the solution. In this case, this means estimated parameters that change infrequently (with the frequency of changes controlled roughly by \(\lambda\)).

We should comment on the difference between using an \(\ell_1\) regularization and some other type of sum-of-norms regularization, such as sum-of-Euclidean norms. With \(\ell_1\) regularization, we obtain a time-varying model in which individual components of the \(\theta(t)\) change infrequently. When we use sum-of-norms regularization, the whole vector \(\theta(t)\) changes infrequently; but when it does change, typically all its components change. In a statistical linear regression framework, sum-of-norms regularization is called Group-Lasso [28], since it results in estimates in which many groups of variables (in this case, all components of the parameter change \(\theta(t) - \theta(t-1)\)) are zero.

2.2 Regularization Path and Critical Parameter Value

The estimated parameter sequence \(\theta(t)\) as a function of the regularization parameter \(\lambda\) is called the regularization path for the problem. Roughly, larger values of \(\lambda\) correspond to estimated \(\theta(t)\) with worse fit, but fewer segments. A basic result from convex analysis tells us that there is a value \(\lambda_{\text{max}}\) for which the solution of the problem is constant, i.e., \(\theta(t)\) does not vary with \(t\), if and only if \(\lambda \geq \lambda_{\text{max}}\). In other words, \(\lambda_{\text{max}}\) gives the threshold above which there is only one segment in \(\theta(t)\). The critical parameter value \(\lambda_{\text{max}}\) is very useful in practice, since it gives a very good starting point in finding a suitable value of \(\lambda\). Reasonable values are typically on the order of \(0.01\lambda_{\text{max}}\) to \(\lambda_{\text{max}}\) (which results in no segmentation).

Let \(\theta^{\text{const}}\) be the optimal constant parameter vector, i.e., the solution of the normal equations
\[\sum_{i=1}^{N} (y_i - \varphi^T(\theta^{\text{const}})) \varphi^T(t) = 0.\]

Then we can express \(\lambda_{\text{max}}\) as
\[\lambda_{\text{max}} = \max_{t=1,\ldots,N-1} \left\{ \sum_{\tau=1}^{t} \| 2(y(\tau) - \varphi^T(\tau)\theta^{\text{const}}) \varphi^T(t) \|_{\text{reg}}^* \right\},\]

where \(\|\cdot\|_{\text{reg}}^*\) is the dual norm associated with \(\|\cdot\|_{\text{reg}}\).

This is readily computed.

To verify our formula for \(\lambda_{\text{max}}\) we use convex analysis [22,4.7]. The constant parameter \(\theta(t) = \theta^{\text{const}}\) solves the problem (6) if and only 0 is in its subdifferential. The fitting term is differentiable, with gradient w.r.t. \(\theta(t)\) equal to
\[2(y(t) - \varphi^T(t)\theta^{\text{const}}) \varphi^T(t).\]

Now we work out the subdifferential of the regularization term. The subdifferential of \(\|\cdot\|_{\text{reg}}\) at 0 is the unit ball in the dual norm \(\|\cdot\|_{\text{reg}}^*\). Therefore the subdifferential of the regularization term is any vector sequence of the
where $z(2), \ldots, z(N)$ satisfy $\|z(t)\|_{\text{reg}} \leq \lambda$. We solve these to get

$$z(t) = -\sum_{\tau=1}^{t-1} g(\tau), \quad t = 2, \ldots, N.$$ 

The optimality condition is

$$g(t) + 2(y(t) - \varphi^T(t)\theta_{\text{const}})\varphi^T(t) = 0, \quad t = 1, \ldots, N.$$ 

Combining this with the formula above yields our condition for optimality of $\theta(t) = \theta_{\text{const}}$ as $\lambda \geq \lambda_{\text{max}}$.

### 2.3 Iterative Refinement

To (possibly) get even fewer changes in the parameter $\theta(t)$, with no or small increase in the fitting term, iterative re-weighting can be used [10]. We replace the regularization term in (6) with

$$\lambda \sum_{t=2}^{N} w(t)\|\theta(t) - \theta(t-1)\|_{\text{reg}},$$

where $w(2), \ldots, w(N)$ are positive weights used to vary the regularization over time. Iterative refinement proceeds as follows. We start with all weights equal to one, i.e., $w^{(0)} = 1$. Then for $i = 0, 1, \ldots$ we carry out the following iteration until convergence (which is typically in just a few steps).

1. **Find the parameter estimate.** Compute the optimal $\theta^{(i)}$ with weighted regularization using weights $w^{(i)}$.
2. **Update the weights.** Set $w^{(i+1)} = 1/(e + \|\theta^{(i)} - \theta^{(i)}(t-1)\|_{\text{reg}})$.

Here $e$ is a positive parameter that sets the maximum weight that can occur.

One final step is also useful. From our final estimate of $\theta(t)$, we simply use the set of times at which a model change occurs (i.e., for which $\theta(t) = \theta(t-1)$ is nonzero), and carry out a final least-squares fit over the parameters, which we now require to be piecewise constant over the fixed intervals. This typically gives a small improvement in fitting, for the same number of segments.

### 2.4 Solution Algorithms and Software

Many standard methods of convex optimization can be used to solve the problem (6) (code used by the authors can be found on [http://www.rt.isy.liu.se/~ohlsson/code.html](http://www.rt.isy.liu.se/~ohlsson/code.html)). Systems such as CVX [13,12] or YALMIP [20] can readily handle the sum-of-norms regularization, by converting the problem to a cone problem and calling a standard interior-point cone solver. For the special case when the $\ell_1$ norm is used as the regularization norm, more efficient special purpose algorithms and software can be used, such as l1_ls [16]. Recently many authors have developed fast first order methods for solving $\ell_1$ regularized problems, and these methods can be extended to handle the sum-of-norms regularization used here; see, for example, [23, §2.2]. Both interior-point and first-order methods have a complexity that scales linearly with $N$.

### 3 Numerical Illustration

We illustrate our method by applying it to a number of segmentation problems. We take $\epsilon = 0.01$ and use the Euclidean norm for regularization throughout the examples. The refinement technique described in Section 2.3 was applied with two refinement iterations and a final refinement by applying least-squares on segments without changes.

#### Example 1 Changing Time Delay

This example is from iddemo11 in the System Identification Toolbox, [19]. Consider the system

$$y(t) + 0.9y(t-1) = u(t-n_k) + e(t).$$

The input $u$ is a $\pm 1$ PRBS (Pseudo-Random Binary Sequence) signal and the additive noise has variance 0.1. At time $t = 20$ the time delay $n_k$ changes from 2 to 1. The data are shown in Figure 1. An ARX-model

$$y(t) + ay(t-1) = b_1u(t-1) + b_2u(t-2)$$

is used to estimate $a$, $b_1$, $b_2$ with the method described in the previous section. The resulting estimates using $\lambda = 0.1\lambda_{\text{max}}$ are shown in Figure 2. The solid lines show the estimate and dashed the true parameter values. We clearly see that $b_1$ jumps from 0 to 1, to “take over” to be the leading term around sample 20. The estimate of the parameter $a$ (correctly) does not change notably.

#### Example 2 Changing Time Series Dynamics

Consider the time series

$$y(t) + ay(t-1) + 0.7y(t-2) = e(t)$$
with $c(t) \sim \mathcal{N}(0, 1)$. At time $t = 100$ the value of $a$ changes from $-1.5$ to $-1.3$. The output data and the estimate of $a$ are shown in Figure 3. $\lambda = 0.01\lambda_{\text{max}}$ was used.

To motivate the iterative refinement procedure suggested in Section 2.3, let us see what happens if it is removed. Figure 4 shows the estimate of $a$ (around $t = 100$) with and without the refinement iteration. As shown by the figure, (6) incorrectly estimates the change at $t = 100$ and gives an estimate having a change both at $t = 100$ and $t = 101$. Using iterative refinement, however, this does not occur. Without iterative refinement, $a$ is estimated to $-5.1$ at $t = 100$.

Example 3  Seismic Signal Segmentation

Let us study the seismic data from the October 17, 1989 Loma Prieta earthquake in the Santa Cruz Mountains. (This data is provided with MATLAB as `quake.mat` and discussed in the command `quake.m`). We choose to decimate the 200 Hz measurements of acceleration in the east-west direction (“e”) by a factor of 100 and segment the resulting signal modeled as an AR process of second order. Here, the regularization constant $\lambda$ in (6) will really act as a design parameter that controls how many segments will be chosen. For example, $\lambda = 0.15\lambda_{\text{max}}$ gives two segments, $\lambda = 0.12\lambda_{\text{max}}$ gives three segments and $\lambda = 0.11\lambda_{\text{max}}$ gives four segments. The result for $\lambda = 0.15\lambda_{\text{max}}$ is shown in Figure 5.
4 Comparisons with Other Methods for Segmentation

Several methods for model segmentation have been suggested earlier, see e.g. [14, Chapter 5], [15], [8]. They typically employ either multiple detection algorithms [24], hidden Markov models (HMM) [5] or explicit management of multiple models, AFMM (adaptive forgetting by multiple models) [2]. The latter algorithm is implemented as the method segment in the System Identification Toolbox and as the routine detectM in the software package adfilt, accompanying the book [15]. The idea is to let $M$ Kalman filters for a stochastic system live in parallel. At each sample the $M$ different predictions from the filters are evaluated. The worst performing filter is killed and a new filter is started. The segmenting is formed by the final estimate of each best performing filter. It should also be mentioned that a similar method to the one proposed in this paper has been discussed for set membership identification, and image segmentation, in [21].

All algorithms for tracking time-varying systems must have a trade-off between assumed noise level ($e$) and the tendency and size of system variations, and that may be reflected in the choice of several tuning parameters. In the segment algorithm, the user has to select 8 parameters (assumed noise variance $R_2$, probability of a jump, the process noise covariance matrix $R_1$, the initial parameter estimates, along with their covariance matrices, the guaranteed life length of each filter, and, if $R_2$ is estimated, the forgetting factor for estimating it). Even though several parameters can be given default values, it may be tedious work to tune the segmented regression algorithm. At the same time it leads to considerable flexibility. For good choices of these parameters, segment often gives performance comparable in quality to the algorithm suggested here. The big advantage of the proposed method is that it has only one scalar design parameter, $\lambda$, with the number of segments controlled by $\lambda$. Moreover, reasonable starting values of the parameter can be found from $\lambda^{\text{max}}$, which is easily computed.

Most existing methods are local in nature: A jump is hypothesized at each time instant, and the ensuing samples are used to test this hypothesis. In contrast, our method is indeed global in nature: For a given $\lambda$ (corresponding to a certain number of jumps), the positions of these jumps are determined as those that globally minimize (6). Still, the complexity of the algorithm is linear in the length of the data record. It seems that this should be an advantage for situations with infrequent jumps in noisy environments. That this indeed is the case is illustrated in the following example.

Example 4 Comparison between segment and (6)

Let us compare our method with segment in the System Identification Toolbox [19]. Consider the system

$$y(t) + a_1 y(t - 1) + 0.7 y(t - 2) = u(t - 1) + 0.5 u(t - 2) + e(t)$$

with $u(t) \sim \mathcal{N}(0, 1)$ and $e(t) \sim \mathcal{N}(0, 9)$. At $t = 400$, $a_1$ changes from $-1.5$ to $-1.3$ and at $t = 1500$ $a_1$ returns to $-1.5$. Both segment and our method are provided with the correct ARX structure and asked to estimate all ARX parameters ($a_1$, $a_2$, $b_1$, $b_2$). With the same design parameters as used to generate the data (the true equation error variance, jump probability, initial ARX parameters and covariance matrix of the parameter jumps) segment does not find any changes at all in the ARX parameters. Tuning the design variable $R_2$ in segment so it finds three segments gives the estimate of $a_1$ shown in Figure 6. It does not seem possible to find values of all the design variables in segment that give the correct jump instants.

Using our method with the same choices as in Section 3 and tuning $\lambda$ so as to obtain three segments gives directly the correct change times. The parameter estimate of our method using $\lambda = 0.025 \lambda^{\text{max}}$ is also shown in Figure 6.
5 Ramifications and Conclusions

5.1 Akaike’s Criterion and Hypothesis Testing

Model segmentation is really a problem of selecting the number of parameters to describe the data. If the ARX model has \( n \) parameters and uses \( R \) segments, the segmented model uses \( d = nR \) parameters. The Akaike criterion (AIC), [1] is a well known way to balance the model fit against the model complexity:

\[
\min_{d \geq 1} \left[ V(Z^N, \Theta) + 2d\sigma^2 \right]
\]

where \( V \) is the negative log likelihood function, \( Z^N \) is the data record with \( N \) observations, and \( \sigma^2 \) is the variance of the innovations. Comparing with (6), \( V \) is the first term (if the innovations are Gaussian), and the regularization term corresponds to the model cardinality term \( 2d\sigma^2 \). In fact, sum-of-norms regularization is a common way to approximate cardinality constraints, e.g. [8]. The link to cardinality penalties becomes even more pronounced with the iterative refinement procedure of Section 2.3. It aims, with iterative replacement of the weights, at a regularization term

\[
\sum_{t=2}^{N} \lambda \left( \| \theta(t) - \theta(t-1) \|_2 + \| \theta(t) - \theta(t-1) \|_1 \right)
\]

which essentially counts the number of nonzero terms, i.e. the number of jumps and hence the number of parameters.

A common statistical approach to selecting model size is to use hypothesis testing, e.g [18, p.507], where the simpler model is the null hypothesis. Using the optimal test, likelihood ratios, is known to correspond to the Akaike criterion at a certain test level, [23]. The criterion (6) can thus be interpreted as a simplified likelihood ratio test, where \( \lambda \) sets the test levels.

5.2 General State Space Models

It is well known that ARX-model estimation with varying parameters can be seen as state estimation in a general state space model, see e.g. [18, p.367]. Applying the Kalman filter to this time-varying ARX-model gives the Recursive Least Squares algorithm. It works well if the time variation is well described as a Gaussian white noise process. The segmentation problem (5) rather correspond to an assumption that the parameter changes at rare instants, i.e. a “process noise” that as zero most of the time, and nonzero at random time instants with a random amplitude. Our method can therefore also be used for state smoothing for general state space models with such process noise. This includes problems of abrupt change detection, and processes with load disturbances (cf equations (2.10)-(2.11) in [18].)

5.3 Summary

We have studied the model segmentation problem and suggested to treat it as least-squares problem with sum-of-norms regularization of the parameter changes. We do not claim that the suggested method necessarily outperforms existing approaches; but being a global method, it certainly has an edge in cases with considerable noise and infrequent jumps. An important benefit is also that it has just one scalar design variable, whose influence on the parameter fit and number of segments is easily understood, and for which a reasonable starting value is readily found.

References


