Automatica ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Brief paper

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

ARTICLE INFO

Article history:
Received 24 July 2009
Received in revised form
3 December 2009
Accepted 10 March 2010
Available online xxxx

Keywords:
Segmentation
Regularization
ARX-models

ABSTRACT

Segmentation of time-varying systems and signals into models whose parameters are piecewise constant in time is an important and well studied problem. Here it is 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.

1. Model segmentation

Estimating linear regression models

$$y(t) = \varphi^T(t)\theta \tag{1}$$

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

$$y(t) + a_1 y(t-1) + \cdots + a_n y(t-n) = b_1 u(t-nk-1) + \cdots + b_m u(t-nk-m) \tag{2}$$

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,

$$\hat{\theta}(N) = \arg\min_{\theta} \sum_{t=1}^{N} \|y(t) - \varphi^T(t)\theta\|^2, \tag{3}$$

where $\|\cdot\|$ denotes the Euclidean or ℓ_2 norm.

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

$$y(t) = \varphi^T(t)\theta(t). \tag{4}$$

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:

$$\theta(t) = \theta_k, \quad t_k < t \le t_{k+1}. \tag{5}$$

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. Bassevill and Nikiforov (1993), Gustafsson (2001), Ljung (1999).

The segmentation problem is often addressed using multiple detection techniques, multiple models and/or Markov models with switching regression, see, e.g. Bodenstein and Praetorius (1977), Lindgren (1978) and Tugnait (1982). The function `segment` for the segmentation problem in the System Identification Toolbox (Ljung, 2007), is based on a multiple model technique (Andersson, 1985).

2. Our method

We shall in this contribution study the segmentation problem from a slightly different perspective. If we allow 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.

☆ The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Er-Wei Bai under the direction of Editor Torsten Söderström. Partially supported by the Swedish foundation for strategic research in the center MOVIII and by the Swedish Research Council in the Linnaeus center CADICS.

* Corresponding author. Tel.: +46 13 282306; fax: +46 13 282622.
E-mail addresses: ohlsson@isy.liu.se (H. Ohlsson), ljung@isy.liu.se (L. Ljung), boyd@stanford.edu (S. Boyd).

0005-1098/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2010.03.013

2.1. Sum-of-norms regularization

To penalize model parameter changes over time, we add a penalty or regularization term (see e.g. Boyd & Vandenberghe, 2004, p. 308) that is a sum of norms of the parameter changes:

$$\min_{\theta(t)} \sum_{t=1}^{N} \|y(t) - \varphi^T(t)\theta(t)\|^2 + \lambda \sum_{t=2}^{N} \|\theta(t) - \theta(t-1)\|_{\text{reg}}, \quad (6)$$

where $\|\cdot\|_{\text{reg}}$ is the norm used for regularization, and λ 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 ℓ_1 or ℓ_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 ℓ_1 norm, i.e., $\|z\|_1 = \sum_{k=1}^{n} |z_k|$, the regularization in (6) is standard ℓ_1 regularization of the least-squares criterion. Such regularization has been very popular recently, e.g. in the much used Lasso method, (Tibsharani, 1996) or compressed sensing (Candès, Romberg, & Tao, 2006; Donoho, 2006). 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 λ, even for large values of N.
- The sum-of-norms form of the regularization favors solutions where "many" (depending on λ) 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 λ).

We should comment on the difference between using an ℓ_1 regularization and some other type of sum-of-norms regularization, such as sum-of-Euclidean norms. With ℓ_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 (Yuan & Lin, 2006), 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 λ is called the *regularization path* for the problem. Roughly, larger values of λ correspond to estimated $\theta(t)$ with worse fit, but fewer segments. A basic result from convex analysis tells us that there is a value $\lambda^{\max}$ for which the solution of the problem is constant, i.e., $\theta(t)$ does not vary with t, if and only if $\lambda \geq \lambda^{\max}$. In other words, $\lambda^{\max}$ gives the threshold above which there is only one segment in $\theta(t)$. The critical parameter value $\lambda^{\max}$ is very useful in practice, since it gives a very good starting point in finding a suitable value of λ. Reasonable values are typically on the order of $0.01\lambda^{\max}$ to $\lambda^{\max}$ (which results in no segmentation).

Let θ^{const} be the optimal *constant* parameter vector, i.e., the solution of the normal equations

$$\sum_{t=1}^{N} (y(t) - \varphi^T(t)\theta^{\text{const}})\varphi^T(t) = 0.$$

Then we can express $\lambda^{\max}$ as

$$\lambda^{\max} = \max_{t=1,\ldots,N-1} \left\| \sum_{\tau=1}^{t} 2(y(\tau) - \varphi^T(\tau)\theta^{\text{const}})\varphi^T(\tau) \right\|_{\text{reg}*}, \quad (7)$$

where $\|\cdot\|_{\text{reg}*}$ is the dual norm associated with $\|\cdot\|_{\text{reg}}$. This is readily computed.

To verify our formula for $\lambda^{\max}$ we use convex analysis (Bertsekas, Nedic, & Ozdaglar, 2003; Borwein & Lewis, 2005; Rockafellar, 1996). 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 form

$$\begin{aligned} g(1) &= -z(2), \\ g(2) &= z(2) - z(3), \\ &\vdots \\ g(N-1) &= z(N-1) - z(N), \\ g(N) &= -z(N), \end{aligned}$$

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 $\lambda \geq \lambda^{\max}$.

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 (Candès, Wakin, & Boyd, 2008). 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., $w^{(0)}(t) = 1$. 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 $\theta^{(i)}(t)$ with weighted regularization using weights $w^{(i)}$.
(2) *Update the weights.*
Set $w^{(i+1)}(t) = 1/(\epsilon + \|\theta^{(i)}(t) - \theta^{(i)}(t-1)\|_{\text{reg}})$.

Here ϵ 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). Systems such as CVX (Grant & Boyd, 2008; Grant, Boyd, & Ye, 2009) or YALMIP (Löfberg, 2004) 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 ℓ_1 norm is used as the regularization norm, more efficient special purpose algorithms and software can be used, such as `l1_ls` (Kim, Koh, Lustig, Boyd, & Gorinevsky, 2007). Recently

Please cite this article in press as: Ohlsson, H., et al. Segmentation of ARX-models using sum-of-norms regularization. Automatica (2010), doi:10.1016/j.automatica.2010.03.013

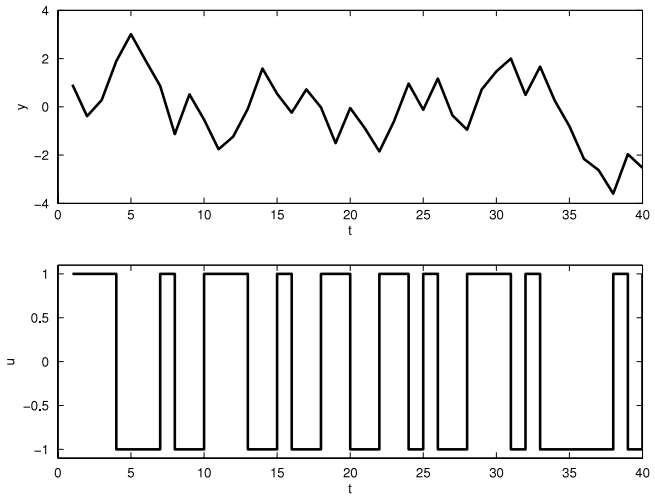

Fig. 1. The data used in Example 1.

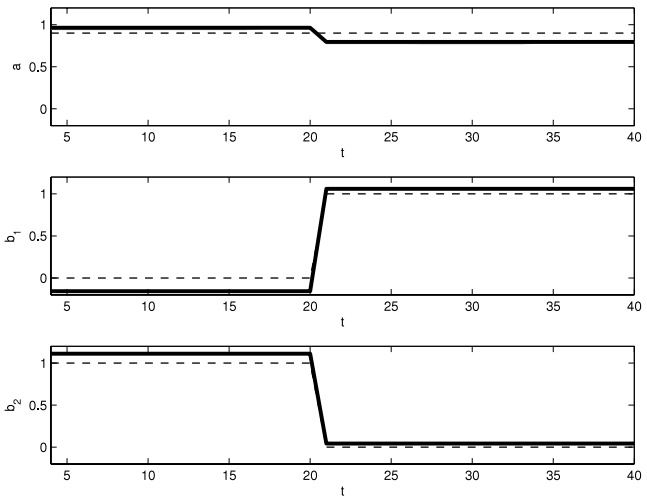

Fig. 2. The parameter estimates in Example 1. Solid lines show the parameter estimates and dashed lines the true parameter values.

many authors have developed fast first-order methods for solving ℓ_1 regularized problems, and these methods can be extended to handle the sum-of-norms regularization used here; see, for example, Roll (2008, Section 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, (Ljung, 2007). Consider the system

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

The input u is a ± 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 Fig. 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 $\lambda = 0.1\lambda^{\max}$ are shown in Fig. 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.

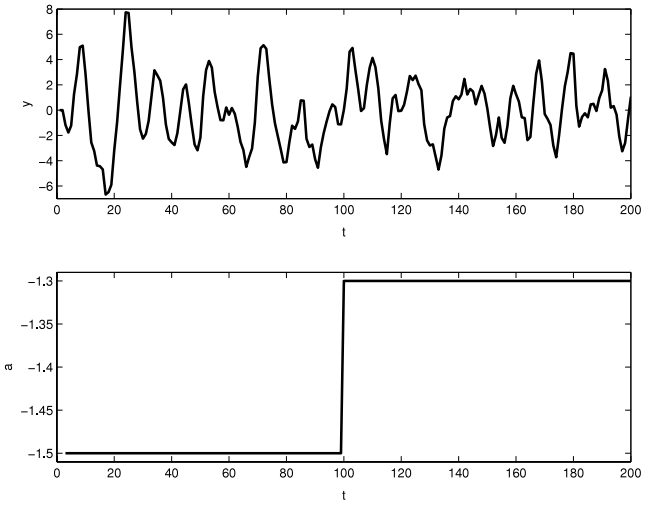

Fig. 3. The time series data (upper plot) and the estimate of a (lower plot) of Example 2.

Example 2 (*Changing Time Series Dynamics*). Consider the time series

$$y(t) + ay(t-1) + 0.7y(t-2) = e(t)$$

with $e(t) \sim \mathcal{N}(0, 1)$. 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 Fig. 3. $\lambda = 0.01\lambda^{\max}$ was used.

To motivate the iterative refinement procedure suggested in Section 2.3, let us see what happens if it is removed. Fig. 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 $t = 100$.

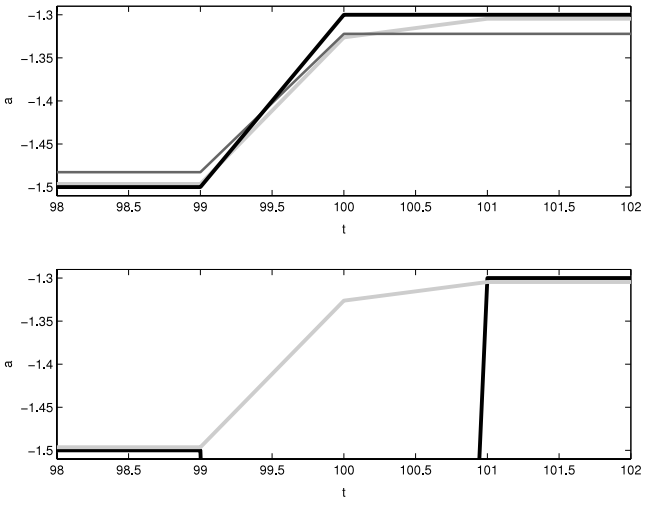

Fig. 4. Estimates of a in Example 2 with (top plot) and without (bottom plot) iterative refinement. Thick black line, estimate after least-squares has been applied to segments without changes in a and light-gray thick line, estimate given by (6). In the top plot, the gray thin lines show estimates of a after one and two iterative refinements (the two lines are not distinguishable). Without iterative refinement (bottom plot) a is estimated to -5.1 at $t = 100$.

Please cite this article in press as: Ohlsson, H., et al. Segmentation of ARX-models using sum-of-norms regularization. Automatica (2010), doi:10.1016/j.automatica.2010.03.013

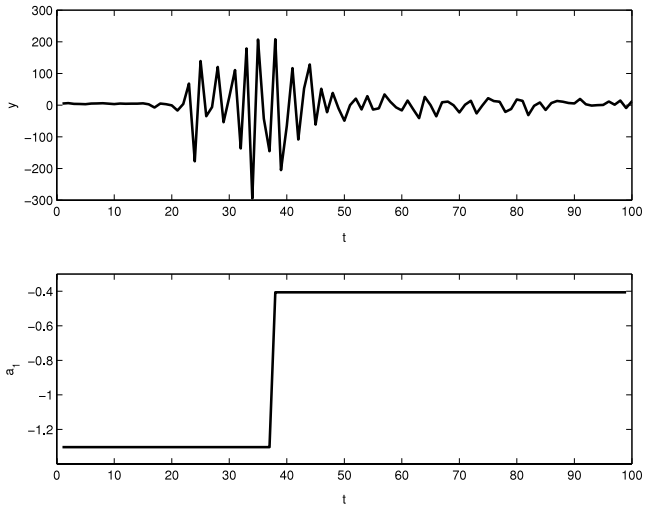

Fig. 5. The seismic signal used in Example 3 is shown in the upper plot. a_1 is shown in the lower plot.

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 λ in (6) will really act as a design parameter that controls how many segments will be chosen. For example, $\lambda = 0.15\lambda^{\max}$ gives two segments, $\lambda = 0.12\lambda^{\max}$ gives three segments and $\lambda = 0.1\lambda^{\max}$ gives four segments. The result for $\lambda = 0.15\lambda^{\max}$ is shown in Fig. 5.

4. Comparisons with other methods for segmentation

Several methods for model segmentation have been suggested earlier, see e.g. Bassevill and Nikiforov (1993), Gustafsson (1992, Chapter 5) and Gustafsson (2001). They typically employ either multiple detection algorithms (Segen & Sanderson, 1980), hidden Markov models (HMM) (Blom & Bar-Shalom, 1988) or explicit management of multiple models, AFMM (adaptive forgetting by multiple models) (Andersson, 1985). 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 (Gustafsson, 2001). 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 segmentation 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 Ozay, Sznaier, Lagoa, and Camps (2008).

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, λ, with the number of segments controlled by λ. Moreover, reasonable starting values of the parameter can be found from $\lambda^{\max}$, which is easily computed.

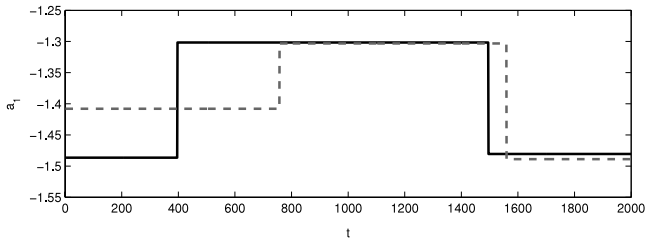

Fig. 6. Estimates of a_1 in the ARX-model used in Example 4 using our method (solid) and `segment` (dashed).

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 λ (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 (Ljung, 2007). Consider the system

$$y(t) + a_1 y(t-1) + 0.7y(t-2) = u(t-1) + 0.5u(t-2) + e(t) \tag{8}$$

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 Fig. 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 λ so as to obtain three segments gives directly the correct change times. The parameter estimate of our method using $\lambda = 0.025\lambda^{\max}$ is also shown in Fig. 6.

5. Ramifications and conclusions

5.1. The Akaike 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 $d = Rn$ parameters. The Akaike criterion (AIC), (Akaike, 1973) is a well known way to balance the model fit against the model complexity:

$$\min_{d,\Theta} \left[V(Z^N, \Theta) + 2d\sigma^2 \right] \tag{9}$$

$$d = \dim(\Theta) \tag{10}$$

where V is the negative log likelihood function, Z^N is the data record with N observations, and σ^2 is the variance of the innovations. Comparing with (6), V is the first term

Please cite this article in press as: Ohlsson, H., et al. Segmentation of ARX-models using sum-of-norms regularization. Automatica (2010), doi:10.1016/j.automatica.2010.03.013

(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. Boyd and Vandenberghe (2004). 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

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

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 Ljung (1999, 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, (Söderström, 1977). The criterion (6) can thus be interpreted as a simplified likelihood ratio test, where λ 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. Ljung (1999, 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 Ljung (1999)).

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

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado.
Andersson, P. (1985). Adaptive forgetting in recursive identification through multiple models. *International Journal of Control*, *42*(5), 1175–1193.
Bassevill, M., & Nikiforov, I. (1993). *Digital signal processing: detection of abrupt changes*. Englewood Cliffs, NJ: Prentice Hall.
Bertsekas, D., Nedic, A., & Ozdaglar, A. (2003). *Convex analysis and optimization*. Athena Scientific.
Blom, H., & Bar-Shalom, Y. (1988). The interacting multiple model algorithm for systems with Markovian switching coefficients. *IEEE Transactions on Automatic Control*, *33*(8), 780–783.
Bodenstein, G., & Praetorius, H. (1977). Feature extraction from the electroencaphalogram by adaptive segmentation. *Proceedings of the IEEE*, *65*, 642–652.
Borwein, J., & Lewis, A. (2005). *Convex analysis and nonlinear optimization: theory and examples*. Canadian Mathematical Society.
Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.
Candès, E., Romberg, J., & Tao, T. (2006). Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, *52*, 489–509.
Candès, E., Wakin, M., & Boyd, S. (2008). Enhancing sparsity by reweighted ℓ_1 minimization. In *Sparsity [Special issue] Journal of Fourier Analysis and Applications*, *14*(5), 877–905.
Donoho, D. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, *52*(4), 1289–1306.
Grant, M., & Boyd, S. (2008). Graph implementations for nonsmooth convex programs. In *Recent advances in learning and control*: *Vol. 371/2008* (pp. 95–110). Berlin, Heidelberg: Springer.
Grant, M., Boyd, S., & Ye, Y. (2009). *CVX: Matlab Software for Disciplined Convex Programming*.
Gustafsson, F. (2001). *Adaptive filtering and change detection*. New York: Wiley.
Gustafsson, F. (1992). Estimation of discrete parameters in linear systems. Ph.D. Linköping University, No. 271.
Kim, S.-J., Koh, K., Lustig, M., Boyd, S., & Gorinevsky, D. (2007). An interior-point method for large-scale l1-regularized least squares. *IEEE Journal of Selected Topics in Signal Processing*, *1*(4), 606–617.
Lindgren, G. (1978). Markov regime models for mixed distributions and switching regressions. *Scandinavian Journal of Statistics*, *5*, 81–91.
Ljung, L. (1999). *System identification-theory for the user* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.
Ljung, L. (2007). *The system identification toolbox: the manual*. Natick, MA, USA: The MathWorks Inc., (1st ed. 1986, 7th ed. 2007).
Löfberg, J. (2004). Yalmip: a toolbox for modeling and optimization in MATLAB. In *Proceedings of the CACSD conference*. Taipei, Taiwan.
Ozay, N., Sznaier, M., Lagoa, C., & Camps, O. (2008). A sparsification approach to set membership identification of a class of affine hybrid systems. In *Proceedings of the 47th IEEE conference on decision and control* (pp. 123–130).
Rockafellar, R. (1996). *Convex analysis*. Princeton University Press.
Roll, J. (2008). Piecewise linear solution paths with application to direct weight optimization. *Automatica*, *44*, 2745–2753.
Segen, J., & Sanderson, A. (1980). Detecting changes in a time-series. *IEEE Transactions on Information Theory*, *26*, 249–255.
Söderström, T. (1977). On model structure testing in system identification. *International Journal of Control*, *26*, 1–18.
Tibsharani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B (Methodological)*, *58*(1), 267–288.
Tugnait, J. (1982). Detection and estimation for abruptly changing systems. *Automatica*, *18*, 607–615.
Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, *68*, 49–67.

Henrik Ohlsson was born in Sweden in 1981. He received the M.Sc. degree in Applied Physics and Electrical Engineering in Oct. 2006 and his Licentiate degree in Automatic Control in Dec. 2008, all from Linköping University, Sweden. He has held visiting positions at the University of Cambridge (UK) and the University of Massachusetts (USA). His research interests are mainly within the areas of system identification and machine learning. He is currently a Ph.D. student at Linköping University.

Lennart Ljung received his Ph.D. in Automatic Control from Lund Institute of Technology in 1974. Since 1976 he is Professor of the chair of Automatic Control In Linkoping, Sweden, and is currently Director of the Strategic Research Center "Modeling, Visualization and Information Integration" (MOVIII). He has held visiting positions at Stanford and MIT and has written several books on System Identification and Estimation. He is an IEEE Fellow, an IFAC Fellow and an IFAC Advisor as well as a member of the Royal Swedish Academy of Sciences (KVA), a member of the Royal Swedish Academy of Engineering Sciences (IVA), an Honorary Member of the Hungarian Academy of Engineering and a Foreign Associate of the US National Academy of Engineering (NAE). He has received honorary doctorates from the Baltic State Technical University in St Petersburg, from Uppsala University, Sweden, from the Technical University of Troyes, France, from the Catholic university of Leuven, Belgium and from Helsinki University of Technology, Finland. In 2002 he received the Quazza Medal from IFAC, in 2003 he received the Hendrik W. Bode Lecture Prize from the IEEE Control Systems Society, and he was the recipient of the IEEE Control Systems Award for 2007.

Stephen Boyd is the Samsung Professor of Engineering in the Electrical Engineering Department at Stanford University. His current interests include convex programming applications in control, signal processing, and circuit design. He received an AB degree in Mathematics, summa cum laude, from Harvard University in 1980, and a Ph.D. in EECS from U.C. Berkeley in 1985. He is the author of Linear Controller Design: Limits of Performance (with Craig Barratt, 1991), Linear Matrix Inequalities in System and Control Theory (with L. El Ghaoui, E. Feron, and V. Balakrishnan, 1994), and Convex Optimization (with Lieven Vandenberghe, 2004).

Please cite this article in press as: Ohlsson, H., et al. Segmentation of ARX-models using sum-of-norms regularization. Automatica (2010), doi:10.1016/j.automatica.2010.03.013