Initialization of Physical Parameter Estimates

Pablo Parrilo, Lennart Ljung

Division of Automatic Control
Department of Electrical Engineering
Linköpings universitet, SE-581 83 Linköping, Sweden
WWW: http://www.control.isy.liu.se
E-mail: ljung@isy.liu.se, @isy.liu.se

3rd December 2003

Report no.: LiTH-ISY-R-2561
Submitted to The 13th IFAC Symposium on System Identification, Rotterdam, The Netherlands

Technical reports from the Control & Communication group in Linköping are available at http://www.control.isy.liu.se/publications.
Abstract: Grey box models of dynamical systems contain designated parameters with physical interpretation to be estimated from input-output data. This often gives distinct advantages over black-box models in terms of fewer parameters to estimate and hence better statistical accuracy. The basic theory for how this can be done is well established. The main practical obstacle may however be how the search for the estimates should be initialized. In this contribution we review the difficulties and point to a possibility to use semidefinite programming and a sum-of-squares formulation to achieve guaranteed consistent initial values for the physical parameters.

1. GREY-BOX MODELS

A linear dynamic model of Grey-box character has the form

\[
\begin{align*}
\dot{x} &= A(\theta)x + B(\theta)u + w \\
y &= C(\theta)x + e
\end{align*}
\] (1)

Here \( y \) and \( u \) are the measured output and input signals, \( x \) are the non-measured states, and \( w \) and \( e \) are non-measurable disturbances. The matrices \( A, B, C \) are matrices with partly known entries, but with some unknown parameters, denoted by \( \theta \). The problem is to estimate these parameters from sampled-data measurements of \( y \) and \( u \).

In this paper we shall assume that there exists a true parameter vector \( \theta_0 \) such that the data have been generated by (1) for this value:

\[
\begin{align*}
\dot{x} &= A(\theta_0)x + B(\theta_0)u + w \\
y &= C(\theta_0)x + e
\end{align*}
\] (2)

The standard solution the linear grey-box problem is to form the Kalman predictor for the sampled data: Equation (1) is transformed to discrete time, in line with the intersample behavior of the input, and then the corresponding Kalman predictor is computed. This will require some assumptions about the noises \( w, e \). These assumptions may or may not be parameterized by \( \theta \). The predictor then takes the form

\[
\begin{align*}
\hat{x}(t + T) &= F(\theta)x(t) + G(\theta)u(t) + K(\theta)(y(t) - C(\theta)\hat{x}(t)) \\
\hat{y}(t|\theta) &= C(\theta)\hat{x}(t)
\end{align*}
\]

Here \( T \) is the sampling interval, and \( F, G, K \) are computed from \( A, B, C, T \) and the noise assumptions in a well known way. For example, if \( w \) is assumed to be 0, then \( K = 0 \), which gives the Output error case.

The parameter vector \( \theta \) is then estimated by minimizing

\[
V_N(\theta) = \sum_{t=1}^{N} ||y(t) - \hat{y}(t|\theta)||^2
\] (3)

with respect to \( \theta \). This is the Prediction error method, closely related to the Maximum likelihood method.

The globally minimizing value to (3) is the estimate, denoted by \( \hat{\theta}_N \). We would like this estimator to be consistent i.e. that \( \hat{\theta}_N \) converges to the true parameter value \( \theta_0 \) as \( N \to \infty \).

In general, the minimization must be done by iterative search of the type
\[ \hat{\theta}_N^{(i+1)} = \hat{\theta}_N^{(i)} + \mu R_N^{(i)} V_N^{(i)} \]  

(4)

See, e.g., (Ljung 1999) for a comprehensive treatment of all this.

The iterative search for the parameter estimate by minimization of (3) is inherent in all Maximum likelihood and prediction error methods for model structures that are not linear regressions (i.e. \( \dot{y}(t|\theta) \) is a non-linear function of \( \theta \)).

We shall in this article only treat the case that the matrices \( A(\theta), B(\theta), C(\theta) \) are such that each entry is either known or a component of \( \theta \). Moreover, no links between the parameters in the different matrices are allowed. This corresponds exactly to the case of the model object \textit{idss} with \textit{SSParameterization = Structured} in the System Identification Toolbox, (Ljung 2000).

The general case, where \( A(\theta), B(\theta), C(\theta) \) are arbitrary functions of \( \theta \) are treated using the model object \textit{idgrey} in the Toolbox, but that case is not treated in this contribution.

Let us remark, that if there are \( p \) outputs, \( m \) inputs and \( n \) states, a maximum of \((p+m)n\) parameters associated with \( A, B, C \) can be estimated. This is the number of parameters in an identifiable canonical representation of the system. So, if the dimension of \( \theta \) is larger than this number, the individual components of the parameter vector cannot be identified, and one can normally as well use a black-box model parameterization. In other words, the process knowledge represented by such a parameterization of (1) has no information value regarding the process dynamics.

2. INITIALIZATION PROBLEMS

The search method (4) requires an initial parameter value \( \hat{\theta}_{N}^{(0)} \) where the search is initialized. Depending on the quality of this initial guess, the iterate \( \hat{\theta}_{N}^{(i)} \) may, as \( i \to \infty \), wander off to infinity, to a local minimum of \( V_N \), get stuck, or converge to the desired value \( \hat{\theta}_N \), i.e. the global minimum of \( V_N \). In other words, we need to find a \( \hat{\theta}_{N}^{(0)} \) that lies in the \textit{domain of attraction} of the global minimum. This is what the current contribution is about.

Is this a pressing problem?

We have generated a number of random, stable systems (using \textit{rss} in MATLAB) and simulated them with noise-free data. We have then randomly picked a certain number of parameters in these models that are to be considered as unknown, to be estimated. We have then minimized \( V_N \) 100 times starting at 100 different randomly chosen initial values of \( \theta \) (but only such values that give stable models), and by Monte Carlo runs estimated the probability of ending up in the correct parameter values.

The results of these simulations are shown in figures 1 – 5.

![Fig. 1. The result of Monte Carlo runs on first order systems. Each star corresponds to a randomly generated system of order 1 with the number of unknown parameters given by the \( x \)-axis. The \( y \)-value is the success-rate in \% for converging to the correct parameters from a random, stable initial model. This number was determined from 100 random initializations. There are 10 stars in each column, each corresponding to a certain system with given unknown parameters.](image1)

![Fig. 2. As figure 1, but for second order systems.](image2)

![Fig. 3. As figure 1, but for third order systems.](image3)
most 5th order system with 10 unknown parameters the probability of convergence to the correct parameter vector from a random initial value is less than 2%. Some remedies that have been suggested include

- Use physical insight for the initialization. Since the parameters have physical significance, it should be possible to have good initial guesses. This is no doubt a very sensible approach. However, the domain of attraction of the global minimum may be small, and prior knowledge may not be sufficient to find it.
- Use global minimization techniques, such as simulated annealing or the genetic algorithm. Conceptually, these can be thought of as (more or less thoughtful) random restarts. The figures which in some cases show 0% success-rate over 100 random initializations give an idea of how difficult this may be.
- Use algebraic tools of different kinds. For example, in (Ljung and Glad 1994) it has been shown that any globally identifiable model structure can be rearranged (at the price of mis-handling the noise structure) to a linear regression. This regression could be used to estimate the initial parameter values using least squares. However, the computational burden in using Ritt’s algorithm for this could be high.

What we seek is a procedure that, asymptotically, as \( N \to \infty \) gives an initial estimate \( \hat{\theta}_N^{(0)} \) that is guaranteed to lie in the domain of attraction of the global estimate \( \hat{\theta}_N \). Since the size of the domain of attraction is not known, this means that this initial estimate \( \hat{\theta}_N^{(0)} \) must itself be a consistent estimate, i.e. \( \hat{\theta}_N^{(0)} \to \theta_0 \) as \( N \to \infty \). The initial estimate is then converted to an estimate \( \hat{\theta}_N \) that is both consistent and efficient (having optimal variance properties) by the minimization scheme (4).

3. USING A BLACK-BOX MODEL FOR INITIALIZATION

What we need is consequently a consistent estimator for \( \hat{\theta} \) in (1), that does not depend on initial guesses. A natural question is if we can employ consistent, non-iterative methods for black-box model for this purpose. A black-box state-space model is given by

\[
\dot{x} = Ax + Bu \\
y = Cx
\]

without any particular structure for the matrices \( A, B \) and \( C \).

So called subspace methods, e.g. (Larimore 1983), (Van Overschee and DeMoor 1996), (Ljung 1999), Chapter 10, have the advantage of producing generically consistent black-box estimates of state space systems, without any iterative search.

Thus a subspace method provides a consistent scheme to produce this system from data, as the data length tends to infinity. In other words, after conversion to continuous time, we will in the limit obtain state space matrices \( A_0, B_0, C_0 \) that realize the same input-output properties as (2).

That means that there exists an invertible matrix \( T \), such that

\[
TA(\theta_0) = A_0T \\
TB(\theta_0) = B_0 \\
C(\theta_0) = C_0T
\]

An obvious question is whether such estimates can be used to find an initial estimate \( \hat{\theta}_0 \) for the minimization of (3).

In some cases this may be straightforward. For example, if the numerators and denominators of \( G(s, \theta) = C(\theta)(sI - A(\theta))^{-1}B(\theta) \) turn out to be linear in the components of \( \theta \), then we can identify coefficients in this expression with

\[
G_0(s) = C_0(sI - A_0)^{-1}B_0
\]

In general, though, identifying such coefficients leads to a set of non-linear equations in \( \theta \). To
solve these numerically may require iterative procedures, and then we are back to the problem of finding a good initial estimate.

4. BASIC SET-UP

The basic approach, as explained in (Xie and Ljung 2002), is to reformulate the question as a polynomial optimization problem, and attempt to solve:

\[
(\hat{\theta}, \hat{T}) = \arg\min_{\theta, T} h(\theta, T) \tag{5a}
\]

\[
h(\theta, T) = \|T \cdot A - A_0(\theta) \cdot T\|_F^2 + \|T \cdot B - B_0(\theta)\|_F^2 + \|C - C_0(\theta)T\|_F^2 \tag{5b}
\]

Here \( \| \cdot \|_F \) denotes the Frobenius norm, i.e. the sum of all squared matrix elements. Under the stated assumptions, the function \( h(\theta, T) \) is a multivariate polynomial in the estimates \( \theta \) and the entries of the matrix \( T \). Our approach to this minimization problem, as outlined in (Parrilo 2000), (Parrilo and Sturmfels 2003), will use sum of squares decompositions to provide a lower bound on the optimal value. If certain additional conditions on the numerical solution are satisfied, we will also obtain the optimal parameter values, corresponding to the global minimizer.

Optimization and sums of squares. We can obtain a lower bound on the optimal value of \( h(\theta, T) \) by finding the largest real number \( \lambda \) for which \( h(\theta, T) - \lambda \) is a sum of squares of polynomials. This can be done by solving a semidefinite program.

We explain the generalities of the method next, referring the reader to the cited works for the details and further applications. The discussion is followed by some specific comments on how to exploit the particular algebraic structure of the polynomial that appears in our problem.

To find a sum of squares decomposition of a multivariate polynomial, we try to express \( h(\theta, T) - \lambda \) as a quadratic form in a properly chosen set of variables \( z \), i.e.,

\[
h(\theta, T) - \lambda = z^T Q z. \tag{6}
\]

The choice of auxiliary variables \( z \) will depend on the structure of the polynomial \( h(\theta, T) \): for instance, for a dense polynomial of degree \( 2d \), we would choose as the \( z_i \) all the monomials of degree \( d \). In general, the variables \( z_i \) will not be algebraically independent, and therefore some quadratic relations (or syzygies) will exist among them. This implies the existence of an affine subspace of matrices \( Q \) for which (6) holds. This subspace will contain a positive semidefinite matrix if and only if the polynomial has a sum of squares representation. By the geometric interpretation of SDP, finding the maximum value of \( \lambda \) for which (6) holds is equivalent to the solution of a semidefinite program.

**Example 1.** Consider the nonconvex quartic polynomial in two variables \( F(\theta, t) \) described below, and define \( z_1 := \theta, z_2 := t, z_3 := t, z_4 = t^2 \):

\[
F(\theta, t) - \lambda = \theta^2 - 4\theta t + 3t^4 - t^2 - \theta + 5 - \lambda
\]

\[
= \begin{bmatrix}
1 & \theta & t \\
\theta & t & t^2
\end{bmatrix}
\begin{bmatrix}
q_{11} & q_{12} & q_{13} & q_{14} \\
q_{12} & q_{22} & q_{23} & q_{24} \\
q_{13} & q_{23} & q_{33} & q_{34} \\
q_{14} & q_{24} & q_{34} & q_{44}
\end{bmatrix}
\begin{bmatrix}
1 \\
\theta \\
t \\
t^2
\end{bmatrix}.
\]

In order for these two expressions to be identical, the following linear equalities should hold:

\[
q_{11} = 5 - \lambda, \quad 2q_{12} = -1, \quad 2q_{13} = 0
\]

\[
q_{33} + 2q_{14} = -1, \quad 2q_{23} = -4, \quad 2q_{24} = 0, \quad 2q_{21} = 1, \quad 2q_{34} = 0, \quad q_{44} = 3.
\]

We want to maximize \( \lambda \) subject to these linear equations, with \( Q \) being positive semidefinite. The optimal \( \lambda, Q \) satisfying all these constraints can be found using semidefinite programming. The optimal solution in this case is given by \( \lambda = 0.75 \) and

\[
Q = \begin{bmatrix}
4.25 & -0.5 & 0 & -3 \\
-0.5 & 1 & -2 & 0 \\
0 & -2 & 5 & 0 \\
-3 & 0 & 0 & 3
\end{bmatrix}.
\]

Factorizing \( Q = L^T L \), we have the sum of squares decomposition:

\[
F(\theta, t) = 0.75 + (0.866025t - 0.866025t)^2 + (0.970143\theta - 2.06155\theta - 0.363803\theta^2)^2 \\
+ (2.06155 - 0.242536\theta - 1.45521t^2)^2.
\]

We have proved thus the lower bound \( F \geq 0.75 \). From the dual solution of the SDP, since it has rank one, we directly obtain the global minimizer \( (\theta, t) = (2.5, 1) \).

Exploiting structure. To achieve efficiency and reliability in the numerical implementation, it is important to exploit as much structure as possible in the polynomial to be minimized. For this, we notice that our objective function \( h(\theta, T) \) in (5b) is a quartic polynomial in the variables \( \theta_i, T_{jk} \). However, since the original equations are bilinear, this implies that the polynomial \( h(\theta, T) \) is actually *biquadratic*, having degree two in each of the \( \theta_i, T_{jk} \) (in other words, no quartic terms in any single variable appear).

As a consequence, in the formulation of the SDP instead of using all the monomials of degree less than or equal to 2 in \( T_{ij}, p_i \) — a total of \( (n^2 + m + 2)/2 \), only those in \( \{1, \theta_i, T_{jk}, \theta_i \cdot T_{jk}\} \) will be needed, with a smaller total of \( (1 + n^2)(1 + \)
m). For instance, if $n = 4, m = 5$, the first number is 253, the second being only 102. In practice however, due to the internal structure of the parameterization, an even more reduced number of parameters may be actually needed. We illustrate these simplifications in Example 2. An important practical remark is that there may be a difference in computational terms between posing the problem as an optimization over $T$ or over $T^{-1}$, since the structure of the polynomial may be simpler in one case versus the other. We also notice that even though nonnegative polynomial are not necessarily sum of squares in the general case, in this specific application the objective function has exactly this form.

Example 2. The physically-based parameterization is given by:

$$A_0 = \begin{bmatrix} 0 & 1 \\ 0 & \theta_1 \end{bmatrix}, \quad B_0 = \begin{bmatrix} 0 \\ \theta_2 \end{bmatrix}, \quad C_0 = \begin{bmatrix} 1 & 0 \end{bmatrix}. \tag{7}$$

The original matrices, estimated using subspace methods, are:

$$A = \begin{bmatrix} -4.125 & -0.111 \\ -1.7815 & -0.114 \end{bmatrix}, \quad B = \begin{bmatrix} 1.3068 \\ 0.4189 \end{bmatrix},$$

$$C = \begin{bmatrix} 0.5518 & -1.669 \end{bmatrix}.$$  

Since there are two uncertain parameters, and the state dimension is two, this corresponds to our notation to the case $n = 2, m = 2$. The procedure described earlier would generate an SDP with matrices of size equal to $15 \times 15$. However, due to the structure of the matrices in (7), only some of the bilinear terms appear. As a consequence, it is possible to further reduce the problem to an $8 \times 8$ semidefinite program indexed by the variables:

$$\theta_1, \theta_2, T_{11}, T_{12}, T_{21}, T_{22}, \theta_1 T_{21}, \theta_1 T_{22}.$$  

After solving the corresponding SDPs, we obtain the optimal parameters and similarity transform:

$$\theta_1 = -4.1871, \quad \theta_2 = 0.9791, \tag{8}$$

$$T = \begin{bmatrix} 0.5405 & -1.6657 \\ 0.7407 & 0.0263 \end{bmatrix}. \tag{9}$$

Results. The method presented works fairly well in small problems. When there are too many parameters and/or states (and here ”too many” is around 20-24 unknowns), the SDP/SOS method slows down considerably due to the large size of the matrices. This is partly a consequence of the use of standard interior-point solvers for solving the corresponding SDPs.

A negative feature of the approach described is that sometimes it may produce answers that, while mathematically correct and optimal, are not useful in terms of the original problem. This can be traced back to the choice of objective function in (5b). Even if we find the globally optimal matrix $T$ that minimizes it (as our method usually does), this may not produce (in theory and/or practice) a ”good” coordinate transformation. The reason for this is that we are not weighting enough the possibility of $T$ being close to singularity. To give a concrete example, for the instance analyzed in (Xie and Ljung 2002), an ”optimal” set of parameters $\theta$ can be obtained: $[0, 0, 0.1639, 0]$. However, the corresponding matrix $T$ achieves a nearly zero objective value, but is very close to singularity with the estimated parameters being far off the “true” ones.

A possibility we are exploring is to modify the objective function given in (5b) by an alternative expression that penalizes directly a measure of the deviation between the models. Natural choices in this regard are system norms such as the 2- or $\infty$-norm of the difference between the corresponding transfer functions, i.e., $\|G(s, \theta) - G_0(s)\|$. In the 2-norm case, for instance, this is a rational function of the unknown parameters. As mentioned earlier, in general this will give rise to a nonconvex problem, but one that we can also attempt to globally minimize using SOS/SDP techniques.

A clear advantage of this approach is that we get rid of all the $T_{ij}$, ending up with a much smaller number of decision variables (equal to the number of parameters $\theta_i$), though the resulting polynomials will in general have higher degree. In this framework, it would also be possible to add additional constraints on the parameters, for instance to restrict them to a range where they are physically meaningful.

5. CONCLUSIONS

We have pointed to one way of initializing the (prediction error) search for physical parameters in structured dynamical models. It gives a solution the problem of consistent initializations and of local minima of the criterion function. The SOS problem gives a consistent estimate of the parameters, as more and more data become available and thus the estimated $A, B, C$ come close, within a similarity transformation to the true system matrices. This consistent estimate will thus approach the domain of attraction of the true parameter values, and the estimate will then have all the asymptotically optimal properties of ML and prediction error estimates.

While the methods as presented work satisfactorily for small problems, further research is needed towards obtaining improved problem formulations, as well as enhancing the computational efficiency of the SOS-based approach.
6. REFERENCES


Abstract

Keywords:
## Abstract

**Title**
Initialization of Physical Parameter Estimates

**Authors**
Pablo Parrilo, Lennart Ljung.