Convexity Issues in System Identification.

Lennart Ljung and Tianshi Chen
Division of Automatic Control, Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden

***Abstract*— System Identification is about estimating models of dynamical systems from measured input-output data. Its traditional foundation is basic statistical techniques, such as maximum likelihood estimation and asymptotic analysis of bias and variance and the like. Maximum likelihood estimation relies on minimization of criterion functions that typically are non-convex, and may cause numerical search problems. Recent interest in identification algorithms has focused on techniques that are centered around convex formulations. This is partly the result of developments in machine learning and statistical learning theory. The development concerns issues of regularization for sparsity and for better tuned bias/variance trade-offs. It also involves the use of subspace methods as well as nuclear norms as proxies to rank constraints. A quite different route to convexity is to use algebraic techniques manipulate the model parameterizations. This article will illustrate all this recent development.**

I. Introduction

System Identification is about building mathematical models of dynamical systems from observed input-output signals.There is a very extensive literature on the subject, with many text books, like [11] and [21]. Most of the techniques for system identification have their origins in estimation paradigms from mathematical statistics, and classical methods like Maximum Likelihood (ML) have been important elements in the area. In this article the main ingredients of this state-of-the-art view of System Identification will be reviewed. This theory is well established and is deployed e.g. in the software [13]. The estimates show attractive asymptotic properties and the methodology has been used extensively and successfully. Some problems can however be listed: (1) the selection of model structures (model orders) is not trivial and may compromise the optimality properties, in particular for shorter data records, and (2) the typically non-convex nature of the criteria may cause numerical optimization artifacts (like ending up in non-global, local minima).

Therefore there is a current trend to enforce estimation methods based on convex formulations. So recently, alternative techniques, mostly from machine learning and the convex optimization area have emerged. Also these have roots in classical statistical (Bayesian) theory. The main elements of these will also be reviewed here.

II. The State-of-the-Art Setup: Parametric Methods

A. Model Structures

A model structure $\mathcal{M}$ is a parameterized collection of models that describe the relations between the input and output signal of the system. The parameters are denoted by θ so $\mathcal{M}(\theta)$ is a particular model. That model gives a rule to predict (one-step-ahead) the output at time t, i.e. $y(t)$, based on observations of previous input-output data up to time $t-1$ (denoted by Z^{t-1}).

$$\hat{y}(t|\theta) = g(t, \theta, Z^{t-1}) \tag{1}$$

For linear systems, a general model structure is given by the transfer function G from input to output and the transfer function H from a white noise source e to output additive disturbances:

$$y(t) = G(q, \theta)u(t) + H(q, \theta)e(t) \tag{2a}$$

$$\mathcal{E}e^2(t) = \lambda; \quad \mathcal{E}e(t)e(k)) = 0 \text{ if } k \neq t \tag{2b}$$

where $\mathcal{E}$ denotes mathematical expectation. This model is in discrete time and q denotes the shift operator $qy(t) = y(t+1)$. We assume for simplicity that the sampling interval is one time unit. For normalization reasons, the function H is supposed to be *monic*, i.e. its expansion starts with a unity. The expansion of $G(q, \theta)$ in the inverse (backwards) shift operator gives the *impulse response* (IR) of the system:

$$G(q, \theta) = \sum_{k=1}^{\infty} g_k(\theta)q^{-k}u(t) = \sum_{k=1}^{\infty} g_k(\theta)u(t-k) \tag{3}$$

The natural predictor for (2a) is

$$\hat{y}(t|\theta) = \frac{H(q, \theta) - 1}{H(q, \theta)}y(t) + \frac{G(q, \theta)}{H(q, \theta)}u(t) \tag{4}$$

Since the expansion of H starts with a "1", the numerator in the first term starts with $h_1 q^{-1}$ so there is a delay in y. The question now is how to parameterize G and H.

1) Black-Box Input-Output Models: Common *black box* (i.e. no physical insight or interpretation) parameterizations are to let G and H be rational in the shift operator:

$$G(q, \theta) = \frac{B(q)}{F(q)}; \quad H(q, \theta) = \frac{C(q)}{D(q)} \tag{5a}$$

$$B(q) = b_1 q^{-1} + b_2 q^{-2} + \ldots b_{nb} q^{-nb} \tag{5b}$$

$$F(q) = 1 + f_1 q^{-1} + \ldots + f_{nf} q^{nf} \tag{5c}$$

$$\theta = [b_1, b_2, \ldots, f_{nf}] \tag{5d}$$

C and D are, like F, monic.

A very common case is that $F = D = A$ and $C = 1$ which gives the *ARX-model*:

$$y(t) = \frac{B(q)}{A(q)}u(t) + \frac{1}{A(q)}e(t) \text{ or} \tag{6a}$$

$$A(q)y(t) = B(q)u(t) + e(t) \text{ or} \tag{6b}$$

$$y(t) + a_1 y(t-1) + \ldots + a_{na} y(t-na) \tag{6c}$$

$$= b_1 u(t-1) + \ldots + b_{nb} u(t-nb) \tag{6d}$$

Other common black/box structures of this kind are FIR (Finite Impulse Response model, $F = C = D = 1$), ARMAX ($F = D = A$), and BJ (Box-Jenkins, all four polynomial different.)

2) Black-box State-Space Models: Another general black-box structure is to use an n:th order state space model

$$x(t+1) = Ax(t) + Bu(t) + Ke(t) \tag{7a}$$

$$y(t) = Cx(t) + e(t) \tag{7b}$$

where the state-vector x is a column vector of dimension n and A, B, C, K are matrices of appropriate dimensions. The parameters θ to estimate consists of all entries of this matrix. Due to possible changes of basis in the state-space, there are many values of θ that correspond to the same system properties. It is easy to see that (7) describes the same models as the ARMAX model with orders n for the A, B, C- polynomials. Also, if the matrix K is fixed to zero, (7) describes the same models as the OE model with orders n for the B, F- polynomials. (See Chapter 4 in [11].)

3) Grey-Box Models: If some physical facts are known about the system, it is possible to build in that into a *Grey-Box Model*. It could, for example be an airplane, for which the motion equations are known from Newton's laws, but certain parameters are unknown, like the aerodynamical derivatives. Then it is natural to build a continuous time state-space models from physical equations:

$$\begin{aligned} \dot{x}(t) &= A(\theta)x(t) + B(\theta)u(t) \\ y(t) &= C(\theta)x(t) + D(\theta)u(t) + v(t) \end{aligned} \tag{8}$$

Here θ corresponds to unknown physical parameters, while the other matrix entries signify known physical behavior. This model can be sampled with the well-known sampling formulas to give

$$\begin{aligned} x(t+1) &= \mathcal{F}(\theta)x(t) + \mathcal{G}(\theta)u(t) \\ y(t) &= C(\theta)x(t) + D(\theta)u(t) + w(t) \end{aligned} \tag{9}$$

See [18] for deeper discussion of sampling of systems with disturbances.

The model (9) has the transfer function from u to y

$$G(q, \theta) = C(\theta)[qI - \mathcal{F}(\theta)]^{-1}\mathcal{G}(\theta) + D(\theta) \tag{10}$$

so we have achieved a particular parameterization of the general linear model (2a).

B. Fitting Time-Domain Data

Suppose now we have collected a data record in the time domain

$$Z^N = \{u(1), y(1), \ldots, u(N), y(N)\} \tag{11}$$

It is most natural to compare the model predicted values (4) with the actual outputs and form the criterion of fit

$$V_N(\theta) = \frac{1}{N}\sum_{t=1}^{N} |y(t) - \hat{y}(t|\theta)|^2 \tag{12}$$

and form the parameter estimate

$$\hat{\theta}_N = \arg\min V_N(\theta) \tag{13}$$

We call this the *Prediction Error Method, PEM*. It coincides with the *Maximum Likelihood, ML,* method if the noise source e is Gaussian. See, e.g. [11] or [16] for more details.

C. Issues of Convexity

For most model structures, the criterion function $V_N(\theta)$ in (12) is non-convex in θ. See for example the plot in Figure 5 in Section VII for an extreme example. That means that the minimization problem (13) requires extra attention, and we can never be quite sure that the global minimum is reached.

It could be noted that ARX-model (6) is an important exception, leading to a quadratic criterion function. According to Section IV these models also have capabilities to approximate arbitrary linear systems. Therefore several estimation algorithms have been developed that capitalize on high-order ARX -models.

A noticeable example is so called *sub-space methods* applied to black-box state-space models (7), e.g. [25]. These methods can be simplistically described as estimating a high-order ARX-model followed by a model reduction step. The essential step in model reduction is to approximate a certain matrix by a lower rank matrix, which is efficiently achieved by a SVD (singular value decomposition) technique. That means that Sub-Space methods are non-iterative estimation methods without issues of local minima. The estimated do not however in general enjoy the same optimal asymptotic properties as the PEM-estimates.

The sub-space algorithms can also be described at fitting IRs to the data, as the same time as keeping the model order small. The model order constraint can be phrased as a certain matrix (Hankel matrix of impulse responses) having a certain rank. Rank constraints may be difficult to handle in efficient algorithms. Therefore it is interesting to relax the rank constraint to a constraint on the nuclear norm (sum of eigenvalues). That is a convexification that has been used successfully e.g. in [9].

D. Asymptotic Properties of the Model estimated by PEM

The observations, certainly of the output from the system are affected by noise and disturbances, which of course also will influence the estimated model (13). The disturbances are typically described as stochastic processes, which makes the

estimate $\hat{\theta}_N$ a *random variable*. This has a certain probability distribution function (pdf) and a mean and a variance.

Except in simple special cases it is quite difficult to compute the pdf of the estimate $\hat{\theta}_N$. However, its *asymptotic properties* as $N \to \infty$ for PEM estimates are easier to establish. The basic results can be summarized as follows: ($\mathcal{E}$ denotes mathematical expectation)

- $$\hat{\theta}_N \to \theta^* = \arg\min \mathcal{E} \lim_{N\to\infty} V_N(\theta) \tag{14}$$

  So the estimate will converge to the best possible model, which gives the smallest average prediction error.

- $$\mathrm{Cov}\hat{\theta}_N \sim \frac{\lambda}{N}\left[\mathrm{Cov}\frac{d}{d\theta}\hat{y}(t|\theta)\right]^{-1} \tag{15}$$

  So the covariance matrix of the parameter estimate is given by the inverse covariance matrix of the gradient of the predictor wrt the parameters. λ is the variance of the optimal prediction errors (*the innovations*). If the model structure contains the true system, it can be shown that this covariance matrix is the smallest that can be achieved by any unbiased estimate. That is, it fulfils the *the Cramér-Rao inequality*, [7]. See [11], chapters 8 and 9 for a general treatment.

These results are valid for quite general model structures. Now, specialize to linear models (2a) and assume that the true system is described by

$$y(t) = G_0(q)u(t) + H_0(q)e(t) \tag{16}$$

which could be general transfer functions, possibly much more complicated than the model. Then we have for the estimated frequency function $G(e^{i\omega}, \hat{\theta}_N)$:

- $$\theta^* = \arg\min_\theta \int_{-\pi}^{\pi} |G(e^{i\omega},\theta) - G_0(e^{i\omega})|^2 \frac{\Phi_u(\omega)}{|H(e^{i\omega},\theta)|^2} d\omega \tag{17}$$

  That is, the frequency function of the limiting model will approximate the true frequency function as well as possible in a frequency norm given by the input spectrum Φ_u and the noise model.

- $$\mathrm{Cov}G(e^{i\omega}, \hat{\theta}_N) \sim \frac{n}{N}\frac{\Phi_v(\omega)}{\Phi_u(\omega)} \text{ as } n, N \to \infty \tag{18}$$

  where n is the model order and Φ_v is the noise spectrum $\lambda|H_0(e^{i\omega})|^2$. The variance of the estimated frequency function at a given frequency is thus, for a high order model proportional to the Noise-to-Signal ratio at that frequency. That is a natural and intuitive result. We see, in particular, that the variance increases with the model order.

III. Bias, Variance and Choice of Model Order

A. Mean Square Error, Bias and Variance

Consider any estimation problem where we estimate a quantity θ. Suppose $\hat{\theta}$ is the estimate and θ_0 is the true value. Denote by $\theta^* = \mathcal{E}\hat{\theta}$ the expected value of the estimate. The difference

$$\theta_B = \hat{\theta} - \theta^* \tag{19}$$

is known as the *bias* of the estimate and it is called *unbiased* if the bias is zero. The *mean square error (MSE)* is

$$\mathcal{E}[(\hat{\theta}_N - \theta_0)(\hat{\theta}_N - \theta_0)^T] = \theta_B\theta_B^T + \mathrm{Cov}\hat{\theta} \tag{20a}$$

$$\mathrm{Cov}\hat{\theta} = \mathcal{E}[(\hat{\theta}_N - \theta^*)(\hat{\theta}_N - \theta^{*T}] \tag{20b}$$

It is thus the sum of bias squared and variance.

B. Trade-off between bias and variance

A goal is really to minimize the MSE with its two contributions: the bias and the variance.

Generally speaking the quality of the model depends on the quality of the measured data and the flexibility of the chosen model structure (1). A more flexible model structure typically has smaller bias, since it is easier to come closer to the true system. At the same time, it will have a higher variance: With higher flexibility it is easier to be fooled by disturbances. (Think of the variance expression (18).) So the trade-off between bias and variance to reach a small total MSE is a choice of balanced flexibility of the model structure.

C. Choice of Model Order

As the model gets more flexible, the fit to the estimation data in (13), $V_N(\hat{\theta}_N)$ will always improve. To account for the variance contribution, it is thus necessary to modify this fit to assess the total quality of the model. A much used technique for this is Akaike's criterion,e.g. [1], which for Gaussian noise e with unknown variance takes the form

$$\hat{\theta}_N = \arg\min\left[\log V_N(\theta) + 2\frac{\dim\theta}{N}\right] \tag{21}$$

were the minimization also takes place over a family of model structures with different number of parameters (dim θ).

A variant of AIC is to put a higher penalty on the model complexity, as in BIC:

$$\hat{\theta}_N = \arg\min\left[\log V_N(\theta) + \log N\frac{\dim\theta}{N}\right] \tag{22}$$

This is known as Akaike's criterion, type B, BIC, or Rissanen's Minimum Description Length (MDL) criterion, [23].

Another important technique is to evaluate the criterion function for the model for another set of data, *validation data*, and pick the model which gives the best fit to this independent data set. This is known as *cross validation*.

As a further element in the choice of model structure various validation criteria should be mentioned. An estimated model should be tried to be *falsified*, i.e. confronted with

facts that may contradict its correctness. A typical fact could be that its residuals (estimated prediction errors) do not show sufficient independence. A good principle is to look for the *simplest unfalsified model*, see e.g.[22].

IV. Approximating Linear Systems by ARX Models

Suppose the true linear system is given by

$$y(t) = G_0(q)u(t) + H_0(q)e(t) \tag{23}$$

Suppose we build an ARX model (6) for larger and larger orders $n = na = nb$:

$$A_n(q)y(t) = B_n(q)u(t) + e(t) \tag{24}$$

Then it is well known from [15] that as the orders tend to infinity at the same time as the number of data N increases even faster we have for the ARX estimate

$$\frac{\hat{B}_n(q)}{\hat{A}_n(q)} \to G_0(q) \tag{25a}$$

$$\frac{1}{\hat{A}_n(q)} \to H_0(q), \qquad \text{as} \quad n \to \infty \tag{25b}$$

This is quite a useful result. ARX-models are easy to estimate. The estimates are calculated by linear least squares techniques, which are convex and numerically robust. Estimating a high order ARX model, possibly followed by some model order reduction could thus be a viable alternative to the numerically more demanding general PEM criterion (13).

The only drawback with high order ARX-models is that they may suffer from high variance. That is the problem we now turn to.

V. Regularization of Linear Regression Models

A. Linear Regressions

A *Linear Regression problem* has the form

$$y(t) = \varphi^T(t)\theta + e(t) \tag{26}$$

Here y (the output) and φ (the regression vector) are observed variables, e is a noise disturbance and θ is the unknown parameter vector. In general $e(t)$ is assumed to be independent of $\varphi(t)$.

It is convenient to rewrite (26) in vector form, by stacking all the elements (rows) in $y(t)$ and $\varphi^T(t)$ to form the vectors (matrices) Y and Φ and obtain

$$Y = \Phi\theta + E \tag{27}$$

The least squares estimate of the parameter θ is

$$\hat{\theta}_N = \arg\min |Y - \Phi\theta|^2 \text{ or} \tag{28a}$$

$$\hat{\theta}_N = R_N^{-1}F_N;\ R_N = \Phi^T\Phi;\ F_N = \Phi^T Y \tag{28b}$$

where $|\cdot|$ is the Euclidean norm.

B. Regularized Least Squares

It can be shown that the variance of $\hat{\theta}$ could be quite large, in particular if Φ has many columns and/or is ill-conditioned. Therefore it makes sense to *regularize* the estimate by a matrix P:

$$\hat{\theta}_N = \arg\min |Y - \Phi\theta|^2 + \theta^T P^{-1}\theta \text{ or} \tag{29a}$$

$$\hat{\theta}_N = (R_N + P^{-1})^{-1}F_N; \tag{29b}$$

The presence of the matrix P will improve the numerical properties of the estimation and decrease the variance of the estimate, at the same time as some bias is introduced. Suppose that the data have been generated by (27) for a certain "true" vector θ_0 with noise with variance $\mathcal{E}EE^T = I$. Then, the mean square error (MSE) of the estimate is

$$\mathcal{E}[(\hat{\theta}_N - \theta_0)(\hat{\theta}_N - \theta_0)^T] = (R_N + P^{-1})^{-1}\times (R_N + P^{-1}\theta_0\theta_0^T P^{-1})(R_N + P^{-1})^{-1} \tag{30}$$

A rational choice of P is one that makes this MSE matrix small. How shall we think of good such choices? It is useful to first establish the following Lemma of algebraic nature

Lemma 1: Consider the matrix

$$M(Q) = (QR + I)^{-1}(QRQ + Z)(RQ + I)^{-1} \tag{31}$$

where I is the identity matrix with suitable dimension, Q, R and Z are positive semidefinite matrices. Then for all Q

$$M(Q) \geq M(Z) \tag{32}$$

where the inequality is in matrix sense.
The proof consists of straightforward calculations, see e.g. [6].

So the question what P gives the best MSE of the regularized estimate has a clear answer: *Use*

$$P = \theta_0\theta_0^T \tag{33}$$

So, not surprisingly the best regularization depends on the unknown system.

We can ask a related question, still from a frequentist perspective: Over a certain set of true systems $\Omega = \{\theta_0\}$ with $\mathcal{E}\theta_0\theta_0 = \Pi$ what is the best *average* MSE? The average MSE is obtained by taking expectation wrt θ_0 over (30). That has the effect that $\theta_0\theta_0^T$ is replaced by Π, so the lemma directly gives the answer:

The best average fit over the set Ω is obtained by the regularization matrix $P = \Pi$.

With this we are very close to a Bayesian interpretation.

C. Bayesian Interpretation

Let us suppose θ is a random vector. That will make y in (27) random variables that are correlated with θ. If the prior (before Y has been observed) covariance matrix of θ is P, then it is known that the maximum a posteriori (after Y has been observed) estimate of θ is given by (29a). [See [6] for all technical details in this section.]

So a natural choice of P is to let it reflect how much is known about the vector θ.

D. "Empirical Bayes"

Can we estimate this matrix P in some way? Consider (27). If θ is a Gaussian random vector with zero mean and covariance matrix P, and E is a random Gaussian vector with zero mean and covariance matrix I, and Φ is a known, deterministic matrix, then from (27) also Y will be a Gaussian random vector with zero mean and covariance matrix

$$Z(P) = \Phi P \Phi^T + I \tag{34}$$

(Two times) the negative logarithm of the probability density function (pdf) of the Gaussian random vector Y will thus be

$$W(Y, P) = Y^T Z(P)^{-1} Y + \log\det Z(P) \tag{35}$$

That will also be the negative log likelihood function for estimating P from observations Y, so the ML estimate of P will be

$$\hat{P} = \arg\min W(Y, P) \tag{36}$$

We have thus lifted the problem of estimating θ to a problem where we estimate parameters (in) P that describe the distribution of θ. Such parameters are commonly known as *hyperparameters*.

If the matrix Φ is not deterministic, but depends on E in such a way that row $\varphi^T(t)$ is independent of the element $e(t)$ in E, it is still true that $W(P)$ in (35) will be the negative log likelihood function for estimating P from Y, although then Y is not necessarily Gaussian itself. [See, e.g. Lemma 5.1 in [11].]

E. FIR Models

Let us now return to the IR (3) and assume it is finite (FIR):

$$G(q, \theta) = \sum_{k=1}^{m} b_k u(t-k) = \varphi_u^T(t)\theta_b \tag{37}$$

where we have collected the m elements of $u(t-k)$ in $\varphi(t)$ and the m IR coefficients b_k in θ_b. That means that the estimation of FIR models is a linear regression problem. All that was said above about linear regressions, regularization and estimation of hyper-parameters can thus be applied to estimation of FIR models. In particular suitable choices of P should reflect what is reasonable to assume about an IR: If the system is stable, b should decay exponentially, and if the IR is smooth, neighboring values should have a positive correlation. That means that a typical regularization matrix P^b for θ_b would be matrix whose k, j element is something like

$$P^b_{k,j}(\alpha) = C\min(\lambda^k, \lambda^j);\ \alpha = [C, \lambda] \tag{38}$$

where $C \geq 0$ and $0 \geq \lambda < 1$. This is one of many possible parameterizations of P (so called *kernels*). This choice is known as the TC-kernel. The hyperparameter α can then be tuned by (36):

$$\hat{\alpha} = \arg\min W(Y, P^b(\alpha)) \tag{39}$$

Efficient numerical implementation of this minimization problem is discussed in [4] and [2]. The routine is implemented as `Impulseest` in the 2012b version of [13].

F. ARX Models

Recall that high order ARX models provide increasingly better approximations of general linear systems. We can write the ARX-model (6) as

$$\begin{aligned} y(t) = &- a_1 y(t-1) - \ldots - a_n y(t-n) + b_1 u(t-1) + \ldots \\ &+ b_m u(t-m) = \varphi_y^T(t)\theta_a + \varphi_u^T(t)\theta_b = \varphi^T(t)\theta \end{aligned} \tag{40}$$

where φ_y and θ_a are made up from y and a in an obvious way. That means that also the ARX model is a linear regression, to which the same ideas of regularization can be applied. Eq (40) shows that the predictor consists of two IRs, one from y and one from u and similar ideas on the parameterization of the regularization matrix can be used. It is natural to partition the P-matrix in (29a) along with θ_a, θ_b and use

$$P(\alpha_1, \alpha_2) = \begin{bmatrix} P^a(\alpha_1) & 0 \\ 0 & P^b(\alpha_2) \end{bmatrix} \tag{41}$$

with $P^{a,b}(\alpha)$ as in (38).

G. Convexity Issues: Multiple Kernel Expressions

In general, the hyper–parameter tuning problem (36) is non-convex, so even if the regularized linear regression problem can be found without problems of local minima, non-convexity shows up in the tuning of the hyperparamters. Since α typically has low dimension, the issue of local minima in solving (39) is less of a problem than that in solving (12) for PEM. Still, it is of interest to consider regularization matrices that are formed linearly from multiple, known kernels P_k:

$$P(\alpha) = \sum_{k=1}^{r} \alpha_k P_k; \quad \alpha_k \geq 0 \tag{42}$$

Actually, this gives several advantages as described in [5], [3], [8]:

- The tuning problem (36) or (39) becomes a *Difference of Convex Functions Programming Problem* which can be solved quite efficiently, [24], [10].
- This kernel offers useful flexibility: The multiple kernel (42) can describe more complex dynamics, e.g., the dynamics contains widely spread time constants and the P_k can be chosen as specific instances of the kernel TC, and also complemented by rank 1 kernels of the kind $\theta_0\theta_0^T$ (cf (33)) for a collection of candidate models θ_0.
- The minimization (39) for (42) favors sparse solutions, i.e. solutions with many $\alpha_k = 0$. That can be very useful if P_k corresponds to different hypothesized structure features in the model. That means that the multiple kernel choice can be applied to a variety of problems in system identification needing sparse solutions.

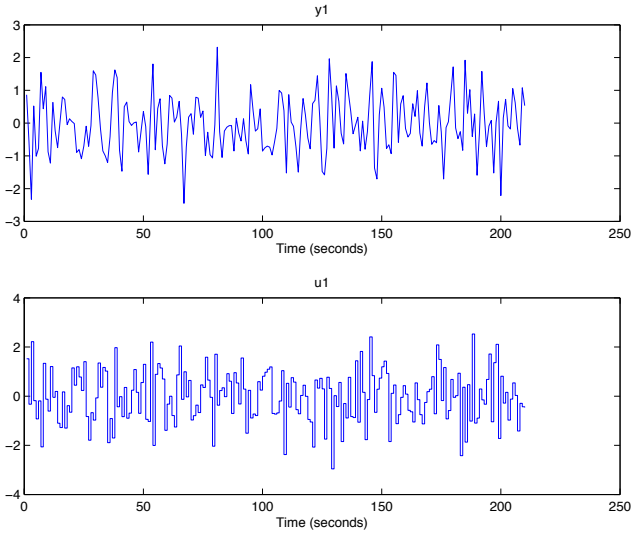

Fig. 1. The investigated data set.

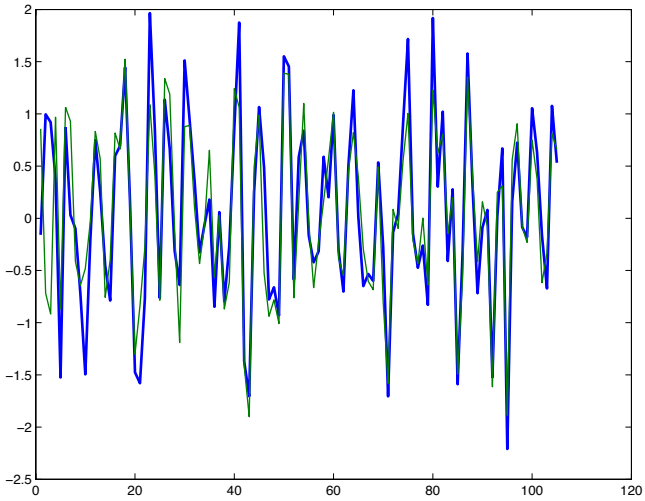

Fig. 2. The measured validation output (thick blue line) together with the simulated output from the model of order 6 (thin green line). The *fit* between the two lines is 61.19%. This "Fit" is the percentage of the output variation that is reproduced by the model.

H. Related work

The text in this section essentially follows [6]. Important contributions of the same kind, based on ideas from machine learning, have been described in [20] and [19]. See also [8]

VI. Numerical Illustration

We will illustrate the issues by estimating models for a particular data set z. It consists of 210 observations from a randomly generated model of order 30. The input is a realization of white Gaussian noise, and white Gaussian noise has also been added to the output of the system, so that the output SNR is about 10. The data set has been selected to illustrate certain points, but the behavior is quite typical for this type of randomly generated high order systems with relatively few observations. The data set is shown in Figure 1.

Now, we want to estimate a model that as accurately as possible reproduces the true system's IR. We only have the data set available and have no knowledge of what is an appropriate order.

A. The state-of-the-art Approach

We will try state-space models (7) with $K = 0$ (equivalent to OE models, according to Section II-A.2) of different orders n. How to choose n?. By cross-validation (see Section III-B) we estimate models using estimation data (typically the first half of the data record) and evaluate how well that model reproduces the other part of the data (the validation data). In terms of MATLAB (The System Identification Toolbox, [13]) we do

```
ze=z(1:105); zv=z(106:210);
mpn=pem(ze,n,'dist','no');
compare(zv,mpn)
```

for different orders n. This gives plots like Figure 2. We try out all orders $n = 1, \ldots, 30$, and the order that gives the best fit for the validation data turns out to be 11 (giving a fit of 61.43%). We could also apply the order selection criteria AIC or BIC (21, 22) to all the 30 tested models. That gives the figures of Table I. The result shows some uncertainty of what order to choose: CV prefers order 11, AIC order 9 and BIC order 6.

Order	Fit	CVFit	AIC	BIC	Actual Fit
1	7.04	-2.14	6.01	4.50	6.89
2	61.28	57.40	58.64	57.30	77.01
4	65.52	60.37	63.52	59.85	**85.80**
6	68.28	61.29	65.46	**60.13**	**83.18**
9	71.19	60.32	**67.26**	59.40	**80.81**
11	71.68	**61.43**	66.88	56.92	**79.57**
17	72.87	56.01	65.40	48.04	77.65
19	72.91	58.07	64.39	43.91	79.66
22	74.00	56.37	64.34	39.67	78.91
29	77.25	-57.89	65.25	30.49	72.61

TABLE I

The fit to (a subset of) models according to different criteria, FIT: The fit to estimation data. CVFit; Fit to validation data as in Figure 2; AIC and BIC: The criteria (21,22), recalculated to a compatible percentage fit; Actual Fit: The fit to the true system according to the oracle.

Now, in this case of simulated data we know the true system and we can let an "oracle" compute the fit between the models' impulse responses and the true one. This actual fit is also inserted in Table I. It shows that the best model of the 30 estimated ones is the model of order 4. This has the most accurate IR, 85.80% of the true IR is reproduced by this 4th order model. See Figure 3 (dashed black curve).

This exercise points to a weakness of the state-of-the-art technique (especially pronounced for short data records): It is not so easy to determine a good model order, or, in other words *to find the best bias-variance trade-off*. The best fit to the true IR we can achieve by state-space models estimated by PEM is *85.80 %, and we might not know that this is the best model* (the CV choice of model gives a fit of 79.57 %).

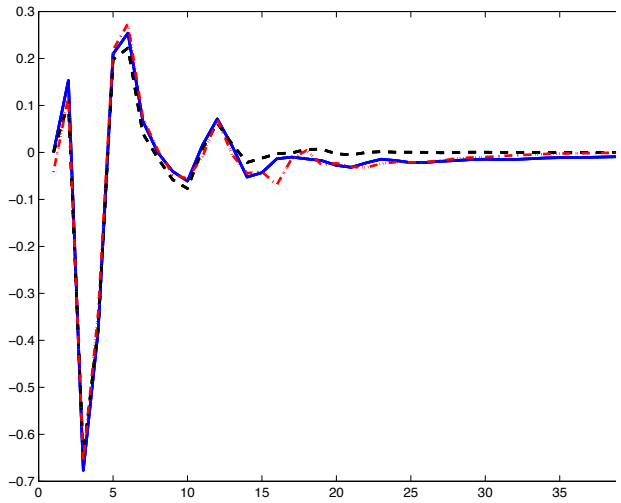

Fig. 3. The true IR (solid,blue), The IR of the best PEM model (dashed black), and the IR of the FIR model (dash-dotted,red)

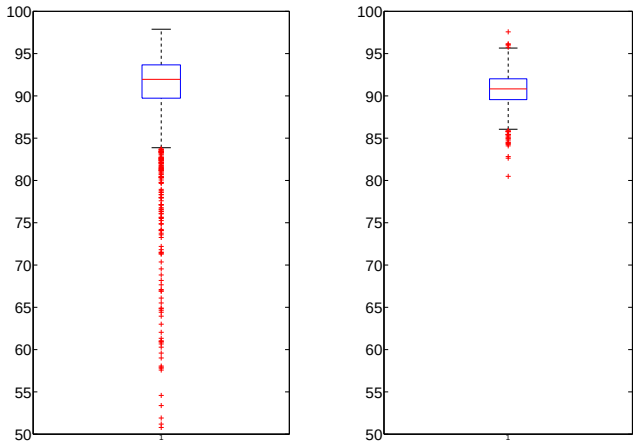

Fig. 4. Box plots over the fits between the IRs of mp10 (left) and mf10 (right) and the true one. 100% means a perfect fit. 2500 systems are shown. [The rectangle in the box-plot shows the 10 –90 percentiles of the fits. The crosses outside the "whiskers" denote outliers.

B. A Regularized FIR Model

Let us now try a regularized FIR model (37) of order 125 with the regularization determined by (38, 36):

```
mf = impulseest(z);
```

We can directly compare the IR (= the FIR coefficients) of this model with the true IR. This is done in Fig 3 (red dash-dotted curve) and we see the fit is **87.51 %**!. *So, the regularized FIR model is better than any of the state space models estimated by PEM.* This shows that *proper bias-variance trade-off is not just picking a suitable model order*. The tuning of the parameters in the regularization matrix (36) can be seen as a continuous moderation that is more flexible than a discrete model order choice.

C. From Regularized FIR to State-space Models

It can be argued against the regularized FIR model `mf` that is is a high order model, that may be more cumbersome to use that the more compact state-space models, so that the comparison is unfair. But it is easy to simplify `mf` to low order (say 6) state-space models by balanced model order reduction:

```
mf6 = balred(mf,6);
```

The IR of `mf6` still shows a 86.87 % fit to the true IR. That can be compared to pem estimate of order 6, `mp6` which has a fit of 83.18 %. *So, we have a 6th order model, mf6, estimated from data, that has a better fit than the 6th order pem estimate from the same data.*

For a particular data set, it is of course no contradiction that specific estimates of a certain order can be better than the PEM estimate of that order. But Monte-Carlo tests over many models and data sets of the same kind show a consistent edge for regularized FIR + balanced model order reduction:

```
mf = impulseest(z);
mf10 = balred(mf,10);
```

```
mp10 = pem(z,10,'dist','no');
```

and comparing the IRs of the true systems with those of `mf10` and `mp10` over many systems shows better accuracy and robustness for the regularized estimates. See Figure 4.

There is no contradiction between these simulations and the (asymptotic) optimality of PEM estimates described in Section II-D. With the small data sets (compared to the high system complexity) we are neither close to the limiting model, nor can we work with accurate model structures that make Cramér-Rao bound relevant. For these data sets the bias-variance trade-off is the most important feature. The regularization offers more flexibility for that trade-off than just mode order selection. In addition, part of the tail in the right boxplot in Figure 4 can be explained by inadequate numerical minimization. (The models have not been subjected to model validation.)

VII. Convexification by Differential Algebra

Let us now illustrate another problem that concerns lack of convexity in system identification loss functions (12). The linear black-box polynomial models can be dealt with by using ARX or FIR models, and black box linear state-space models can be dealt with by subspace methods.

A more serious problem is when non-linear grey-box models show severe non-linearities. We shall look at a variant of the Michaelis-Menten growth kinetic equations which describe the growth of enzyms materia.

If we denote by y the concentration of a certain enzyme, and by u the addition of nutrition substrate, the dynamics is described by

$$\dot{y} = \theta_1 \frac{y}{\theta_2 + y} - y + u \tag{43}$$

Here θ_1 and θ_2 are the maximal growth rate and the Michaelis constant respectively, which are specific for a cer-

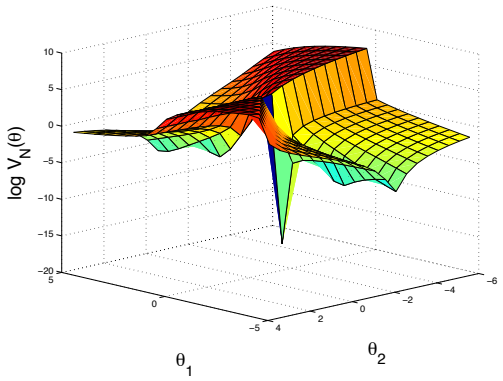

Fig. 5. The ML criterion $V_N(\theta)$ for the model parameterization

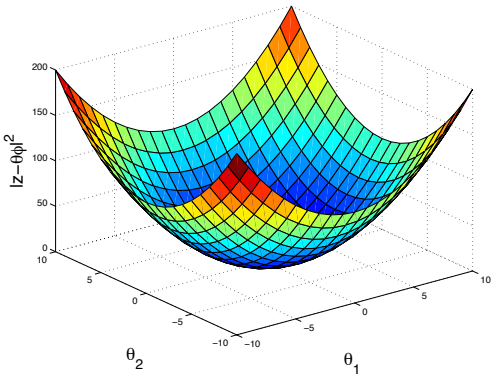

Fig. 6. The surface for reorganized equations relating the parameters to the misfit $\|z - \theta\phi\|^2$

tain enzyme. We assume that we measure the concentration at time t_k with a certain measurement error $e(k)$:

$$y_m(t_k) = y(t_k) + e(k) \tag{44}$$

The predictor (1) and the criterion function (12) can readily be found as

$$V_N(\theta) = \sum_{k=1}^{N} [y_m(t_k) - \hat{y}(t_k|\theta)]^2 \tag{45}$$

$$\dot{\hat{y}}(t|\theta) = \theta_1 \frac{\hat{y}(t|\theta)}{\theta_2 + \hat{y}(t|\theta)} - \hat{y}(t|\theta) + u(t) \tag{46}$$

Assume $u(t)$ is an impulse at time 0. Note that the shape of the loss function is independent on the noise level (size of the variance of e.) It is thus sufficient to plot (45) for $e \equiv 0$.That plot is shown in Figure 5. It is clearly seen that even for noise-less observations it is quite a challenging task to find $\arg\min V_N(\theta)$. Very good initial guesses are required to have success with iterative search. This is true no matter how small the variance of the noise is. One might conclude that it is difficult to find the parameters of this model, and that information about them are well hidden in the data.

If we for the moment disregard the noise e, we can do as follows: Multiply (43) with the numerator and rearrange the terms:

$$\dot{y}y + \theta_2\dot{y} = \theta_1 y - y^2 - \theta_2 y + u + \theta_2 u$$

or

$$\dot{y}y + y^2 - uy = \begin{bmatrix}\theta_1 & \theta_2\end{bmatrix} \begin{bmatrix} y \\ u - \dot{y} - y \end{bmatrix} \tag{47a}$$

$$\text{or } z = \theta^T \phi \tag{47b}$$

with obvious definitions of z and ϕ. Equation (47b) is a linear regression that relates the unknown parameters and measured variables ϕ and z. We can thus find them by a simple least squares procedure. See Figure 6.

The manipulations leading to (47a) are an example of Ritt's algorithm in Differential Algebra. In fact it can be shown, [14], that any *globally identifiable model structure can be rearranged* (using Ritt's algorithm) *to a linear regression*. This is in a sense a general convexification result for any identifiable estimation problem. A number of cautions must be mentioned, though:

- Although Ritt's algorithm is known to converge in a finite number of steps, the complexity of the calculations may be forbidding for larger problems.
- With noisy measurements, care must be exercised in differentiation, and also the linear regression may be subject to disturbances that can give biased estimates.

But the fact remains: the result shows that the complex, non-convex form of the likelihood function with many local minima is not inherent in the model structure.

VIII. Conclusions

System Identification is an area of clear importance for practical systems work. It has now a well developed theory and is a standard tool in industrial applications. Even though the area is quite mature with many links to classical theory, new exciting and fruitful ideas keep being developed. This article has tried to focus on some current work that has convexification as a prime goal. Further discussions and views on the current status and future perspectives on system identification are given in e.g. [12] and [17].

IX. Acknowledgements

This work was supported by the ERC advanced grant LEARN, under contract 267381.

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19:716–723, 1974.

[2] F.P. Carli, A. Chiuso, and G. Pillonetto. Efficient algorithms for large scale linear system identification using stable spline estimators. In *Proceedings of the 16th IFAC Symposium on System Identification (SysId 2012)*, 2012.

[3] Tianshi Chen, Martin S. Andersen, Lennart Ljung, Alessandro Chiuso, and Gianluigi Pillonetto. System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques. *IEEE Transactons on Automatic Control*, Submitted, Oct. 29 2012.

[4] Tianshi Chen and Lennart Ljung. Implementation of algorithms for tuning parameters in regularized least squares problems in system identification. *Automatica*, 50, 2013. to appear.

[5] Tianshi Chen, Lennart Ljung, Martin Andersen, Alessandro Chiuso, P. Carli Francesca, and Gianluigi Pillonetto. Sparse multiple kernels for impulse response estimation with majorization minimization algorithms. In *IEEE Conference on Decision and Control*, pages 1500–1505, Hawaii, Dec 2012.

[6] Tianshi Chen, Henrik Ohlsson, and Lennart Ljung. On the estimation of transfer functions, regularizations and Gaussian processes-Revisited. *Automatica*, 48(8):1525–1535, 2012.

[7] H. Cramér. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, N.J., 1946.

[8] F. Dinuzzo. Kernels for linear time invariant system identification. Manuscript, Max Planck Institute for Intelligent Systems, Spemannstrasse 38,72076 Tübingen, Germany, 2012.

[9] C. Grossman, C. N. Jones, and M.Morari. System identification via nuclear norm regularization for simulated moving bed processes from incomplete data sets. In *Proc. IEEE Conference on Decision and Control, 2009*, Shanghai, China, Dec.

[10] R. Horst and N. V. Thoai. DC programming: Overview. *Journal of Optimization Theory and Applications*, 103(1):1–43, 1999.

[11] L. Ljung. *System Identification - Theory for the User*. Prentice-Hall, Upper Saddle River, N.J., 2nd edition, 1999.

[12] L. Ljung. Pespectives on system identification. *IFAC Annual Reviews*, Spring Issue, 2010.

[13] L. Ljung. *The System Identification Toolbox: The Manual*. The MathWorks Inc. 1st edition 1986, 8th edition 2012, Natick, MA, USA, 2012.

[14] L. Ljung and T. Glad. On global identifiability of arbitrary model parameterizations. *Automatica*, 30(2):pp 265–276, Feb 1994.

[15] L. Ljung and B. Wahlberg. Asymptotic properties of the least-squares method for estimating transfer functions and disturbance spectra. *Adv. Appl. Prob.*, 24:412–440, 1992.

[16] Lennart Ljung. Prediction error estimation methods. *Circuits, systems, and signal processing*, 21(1):11–21, 2002.

[17] Lennart Ljung, Hakan Hjalmarsson, and Henrik Ohlsson. Four Encounters with System Identification. *European Journal of Control*, 17(5-6):449–471, 2011.

[18] Lennart Ljung and Adrian Wills. Issues in sampling and estimating continuous-time models with stochastic disturbances. *AUTOMATICA*, 46(5):925–931, 2010.

[19] G. Pillonetto, A. Chiuso, and G. De Nicolao. Prediction error identification of linear systems: a nonparametric Gaussian regression approach. *Automatica*, 47(2):291–305, 2011.

[20] G. Pillonetto and G. De Nicolao. A new kernel-based approach for linear system identification. *Automatica*, 46(1):81–93, January 2010.

[21] R. Pintelon and J. Schoukens. *System Identification – A Frequency Domain Approach*. IEEE Press, New York, 2nd edition, 2012.

[22] K. R. Popper. *The Logic of Scientific Discovery*. Basic Books, New York, 1934.

[23] J. Rissanen. Modelling by shortest data description. *Automatica*, 14:465–471, 1978.

[24] P. D. Tao and L. T. H. An. Convex analysis approach to D. C. programming: Theory, Algorithms and Applications. *ACTA Mathematica Vietnamica*, 22:289–355, 1997.

[25] P. Van Overschee and B. DeMoor. *Subspace Identification of Linear Systems: Theory, Implementation, Applications*. Kluwer Academic Publishers, 1996.