Improved initial approximation for errors-in-variables system identification

Konstantin Usevich

***Abstract*— Errors-in-variables system identification can be posed and solved as a Hankel structured low-rank approximation problem. In this paper different estimates based on suboptimal low-rank approximations are proposed. The estimates are shown to have almost the same efficiency and lead to the same minimum when supplied as an initial approximation for local optimization in the structured low-rank approximation problem. In this paper it is shown that increasing Hankel matrix window length improves initial approximation for autonomous systems and does not improve it in general for systems with inputs.**

I. INTRODUCTION

The problem of system identification for linear time-invariant systems where both inputs and outputs are subject to measurements errors has recently received much attention. For a survey of methods see, for example [12].

One of possible approaches to errors-in-variables system identification is to formulate it as a Hankel structured low-rank approximation problem (SLRA, [13], [17]). SLRA solution can be interpreted as a maximum likelihood estimator of the true trajectory under assumption of Gaussian errors. Moreover, SLRA provides a consistent estimator of the system under weaker assumptions of zero-mean errors with covariance structure known up to a scalar factor [10].

SLRA is a nonconvex optimization problem and has no unique solution. There are two main ways of solving SLRA problem: local optimization methods and heuristics, which are based on relaxations of the problem. It was shown previously for scalar autonomous systems that the accuracy of suboptimal subspace methods can be improved if one takes bigger window length of the Hankel matrix [14]. In this paper compare different subspace-based methods as initial approximations for SLRA local optimization.

In Section II we formulate the Hankel structured low-rank approximation problem, establish links to the errors-in-variables identification. In Section III we propose different initial approximations based on suboptimal solutions of the SLRA problem for the autonomous case and the case of systems with inputs. In Section IV we perform numerical experiments for different initial approximations.

II. HANKEL STRUCTURED LOW-RANK APPROXIMATION

A. Problem formulation and kernel parametrisation

Let $w = [w(1), \dots, w(T)] \in (\mathbb{R}^q)^{\mathrm{T}}$ be a finite time series. A $qL \times (T-L+1)$ block Hankel matrix, parametrised by the time series and a *window length* L, is defined as

$$\mathscr{H}_L(w) = \begin{bmatrix} w(1) & w(2) & \cdots & w(T-L+1) \\ w(2) & w(3) & \cdots & w(T-L+2) \\ \vdots & \vdots & \vdots & \vdots \\ w(L) & \cdots & \cdots & w(T) \end{bmatrix}.$$

The Hankel structured low-rank approximation problem is defined as follows.

Problem 1: Given $w_d \in (\mathbb{R}^q)^{\mathrm{T}}$, L and $r < qL$

$$\underset{\widehat{w} \in (\mathbb{R}^q)^{\mathrm{T}}}{\text{minimize}} \ \|\widehat{w} - w_d\|_2 \ \text{ subject to } \ \operatorname{rank} \mathscr{H}_L(\widehat{w}) \leq r. \quad (1)$$

For a time series w the rank constraint $\mathscr{H}_L(w) \leq r$ is equivalent to

$$R\mathscr{H}_L(w) = 0, \quad (2)$$

where $R \in \mathbb{R}^{p \times qL}$ is a full row rank matrix and $p := qL - r$ is the rank reduction. We can rewrite (2) as a recurrent relation

$$R_0 w(t) + R_1 w(t+1) + \cdots + R_\ell w(t+\ell) = 0, \quad (3)$$

where $\ell = L - 1$, $R = \begin{bmatrix} R_0 & \dots & R_\ell \end{bmatrix}$, $R_k \in \mathbb{R}^{p \times q}$ and $1 \leq t \leq T - \ell$.

B. Linear-time-invariant systems

We use behavioural system-theoretic terminology (see [2] or [11, Ch. 7]) as it is more convenient for formulation of errors-in-variables identification problem, where inputs and outputs are treated on an equal footing.

A discrete-time *dynamical system* $\mathscr{B}$ with q variables is a subset of $(\mathbb{R}^q)^{\mathbb{N}}$. A dynamical system $\mathscr{B}$ is *linear-time-invariant (LTI)* if there exists a nontrivial sequence $R_0, \dots, R_l \in \mathbb{R}^{p \times q}$ (a *kernel representation* of the system) such that $\mathscr{B}$ is characterised by (3), i.e.

$$w \in \mathscr{B} \iff (3) \text{ holds } \forall t \in \mathbb{N}.$$

An LTI system can possess many kernel representations. The minimal among kernel representations $\ell(\mathscr{B}) := \ell_{min}$ and $\mathbf{p}(\mathscr{B}) = p_{min}$ are called the *lag* and the *output cardinality* of the system respectively. The latter has the usual meaning in classical systems theory, i.e. there exists a partition

$$w(t) = \Pi(u(t), y(t)), \quad y(t) \in \mathbb{R}^{\mathbf{p}(\mathscr{B})}, y(t) \in \mathbb{R}^{\mathbf{m}(\mathscr{B})}, \quad (4)$$

where $\mathbf{m}(\mathscr{B}) := q - \mathbf{p}(\mathscr{B})$ is the *input cardinality*, such that the system admits an input/state/output representation

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t), \end{aligned} \quad (5)$$

where $x(t) \in (\mathbb{R}^n)^{\mathbb{N}}$ is the *state variable* and A, B, C, D are real matrices of appropriate dimensions. The system is

K. Usevich is a Research Fellow at School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, United Kingdom kdu@ecs.soton.ac.uk

called autonomous if $\mathbf{m}(\mathscr{B}) = q$ (or all the variables are outputs). The *minimal state dimension* $\mathbf{n}(\mathscr{B})$ does not depend on input/output partition and is bounded $\mathbf{p}(\mathscr{B})\ell(\mathscr{B})$. For more details see, for example, [11, Ch. 7].

C. Errors-in-variables system identification

Let $\mathscr{L}^q_{m,\ell}$ denote the model class of LTI systems with q variables, at most m inputs and lag at most ℓ. Let $\mathscr{B}|_{[1,T]} \subset (\mathbb{R}^q)^T$ denote the set of finite trajectories of $\mathscr{B}$. The errors-in-variables system identification problem may be formulated as an approximation problem or estimation problem [13].

Problem 2: Given a finite trajectory $w_d \in (\mathbb{R}^q)^T$ generated as

$$w_d = w_0 + \widetilde{w}, w_0 \in \mathscr{L}^q_{m,\ell}|_{[1,T]} \tag{6}$$

- *(approximation problem)*
  find the best approximation $\widehat{w} \in \mathscr{L}^q_{m,\ell}|_{[1,T]}$ of w_d.
- *(estimation problem)*
  find an estimate $\widehat{w} \in \mathscr{L}^q_{m,\ell}|_{[1,T]}$ of the true trajectory w_0 in the model (6) with additional assumptions on distribution of $\widetilde{w}$.

The next proposition is a corollary of [11, Theorem 11.4].

Proposition 1: Let (3) be a kernel representation of $\mathscr{B} \in \mathscr{L}^q_{m,\ell}$, and $\operatorname{rank} R_\ell = q - m$. Then for any $w \in (\mathbb{R}^m)^T$

$$w \in \mathscr{B}|_{[1,T]} \iff \begin{bmatrix} R_0 & \dots & R_l \end{bmatrix} \mathscr{H}_{l+1}(w) = 0.$$

Proposition 1 establishes a link between Hankel low-rank approximation and errors-in-variables system identification.

Corollary 1: Problem 2 with 2-norm approximation criterion is equivalent to Problem 1 with $L = \ell+1$, rank reduction equal to the number of outputs $p = q - m$ and additional constraints on the kernel of the Hankel matrix.

Corollary 1 shows that structured low-rank approximation provides a natural way of errors-in-variables system identification in behavioural setting, without choosing a specific input/output partition.

D. Variable projection solution of the SLRA problem

Application of the variable projection principle [16], [17] to Problem 1 leads to the following optimization problem.

Problem 3:

$$\underset{R:\, \operatorname{rank} R = p}{\text{minimize}} \; f(R), \quad \text{where}$$

$$f(R) = \min_{\widehat{w}} \|\widehat{w} - w\|_2 \;\; \text{subject to} \;\; R\mathscr{H}_L(\widehat{w}) = 0. \tag{7}$$

Variable projection has the following features that make it attractive in context of errors-in-variables identification:

- the inner minimisation problem (7) is a least-norm problem, and can be solved efficiently due to the structure of the problem;
- number of optimization parameters is small compared to the original problem;
- the optimization variable R, by Proposition 1, is related to kernel representation, and the minimum gives an estimate of the kernel representation of the true system.

We will consider even more restricted problem

$$\underset{X:\, X \in \mathbb{R}^{p \times Lq}}{\text{minimize}} \; f\left(\begin{bmatrix} X & -I_p \end{bmatrix}\right), \tag{8}$$

which is a *structured total-least squares* (STLS) problem [11, Ch. 4].

Since $f(R)$ is a function of row subspace of the matrix R [16], the problem (8) is a restriction of Problem 3 to the set of matrices of form $R = S\begin{bmatrix} X & -I_p \end{bmatrix}$ (S is a nonsingular $p \times p$ matrix), which is a generic subset of the set the matrices R satisfying conditions of Proposition 1 ([11, Ch.11]). This restriction is also equivalent to existence of input/output partition with $\Pi = I$.

We use restricted problem (8) its solution is a consistent estimator [10] of w_0 Problem 2 under assumption of zero-mean errors with covariance structure known up to a scalar factor (and some additional technical assumptions). This type of estimator is also related to Markov estimator for semilinear models [6].

III. SUBOPTIMAL SUBSPACE METHODS

The base of all subspace methods considered in this section is the following relaxation of the structured low-rank approximation problem.

Problem 4: Given w_d, L and $r < qL$

$$\underset{\widehat{H} \in \mathbb{R}^{qL \times (T-L+1)}}{\text{minimize}} \; \|\widehat{H} - \mathscr{H}_L(w_d)\|_F \;\; \text{subject to} \;\; \operatorname{rank} \widehat{H} \leq r.$$

The solution of this problem is a truncated singular value decomposition (SVD)

$$\widehat{H}^{(LRA)}_{r,L}(w_d) = \sum_{k=1}^{r} \sigma_k U_k V_k^{\mathrm{T}}, \tag{9}$$

where (σ_k, U_k, V_k) are *singular triplets* of $\mathscr{H}_L(w_d)$ [17].

A. Default (TLS) initial approximation

If one takes $L = \ell + 1$ in (9), then one obtains the default initial approximation for SLRA local optimization.

Algorithm 1 (LRA initial approximation for SLRA):

1) Compute the SVD $\mathscr{H}_{\ell+1}(w_d) = \sum_{k=1}^{(\ell+1)q} \sigma_k U_k V_k^{\mathrm{T}}$.
2) Set

$$\begin{aligned} \widehat{R}^{(0)}_{m,\ell}(w_d) = \widehat{R}^{(0)}_{m,\ell}(\mathscr{H}_{\ell+1}(w_d)) := \\ \begin{bmatrix} U_{\ell q+m+1} & \dots & U_{(\ell+1)q} \end{bmatrix}^{\mathrm{T}}. \end{aligned} \tag{10}$$

Remark 1: Initial approximation for STLS problem is computed as

$$\widehat{X}(\widehat{R}) := Q^{-1}P,$$

where $\begin{bmatrix} P & Q \end{bmatrix}^{\mathrm{T}} := \widehat{R}$.

It is known that in subspace methods L is often taken to be larger than the system lag ℓ, in order to achieve better signal/noise separation. For example, in [7] it is shown for scalar autonomous systems that increasing L leads to better estimates of the parameters of the system. In what follows, we will consider different subspace-based methods in view of initial approximation for Hankel structured low-rank approximation.

B. Diagonal averaging and Cadzow iterations

Suppose one aims to estimate the original signal w_0 in the sum (6). The matrix $\widehat{H}_{r,L}^{(LRA)}(w_d)$ obtained by low-rank approximation is not a block Hankel matrix, but one could enforce the block-Hankel structure by computing a diagonal averaged time series $\widehat{w}^{(diag)} = \widehat{w}^{(diag)}(\widehat{H})$:

$$(\widehat{w}^{(diag)}(t))_j := \frac{\sum\limits_{l+k-1=t} \widehat{H}_{q(l-1)+j,k}}{\min\{t, L, N-t+1\}}.$$

Diagonal averaging is nothing more but orthogonal projection on the space of (block) Hankel matrices. The estimate

$$\widehat{w}_{r,L}^{(SSA)} := \widehat{w}^{(diag)}(\widehat{H}_{r,L}^{(LRA)}(\mathscr{H}_L(w_d)))$$

corresponds to Singular Spectrum Analysis *reconstruction* of w_0 [7]. Note that $\mathscr{H}_L(\widehat{w}_{r,L}^{(SSA)})$ is not of low rank, but $\widehat{w}_{r,L}^{(SSA)}$ can serve as a pointwise estimate of values of w_0.

This procedure also corresponds to Cadzow iterations.

Proposition 2 ([3]): Define $\widehat{w}_{r,L,1}^{(Cad)} := \widehat{w}_{r,L}^{(SSA)}(w)$ and $\widehat{w}_{r,L,k+1}^{(Cad)} := \widehat{w}_{r,L}^{(SSA)}(\widehat{w}_{r,L,k}^{(Cad)})$. Then the $\widehat{w}_{r,L,k}^{(Cad)} \to \widehat{w}_{r,L,\infty}^{(Cad)}$ as $k \to \infty$, and

$$\operatorname{rank} \mathscr{H}_L(\widehat{w}_{r,L,\infty}^{(Cad)}) \le r. \tag{11}$$

Cadzow iteration process gives a suboptimal solution of the structured low-rank approximation problem.

C. Proposed improved initial approximations for SLRA

In this section we propose several initial approximations for kernel in SLRA optimization that exploit only the following rank property of the Hankel matrix.

Lemma 1 ([11, Ch.7]): For an LTI system $\mathscr{B}$ and $L \ge \ell(\mathscr{B})$ we have

$$\begin{aligned}\dim(\mathscr{B}|_{[1,L]}) = \mathbf{m}(\mathscr{B})L + \mathbf{n}(\mathscr{B}) \le \\ \le \mathbf{m}(\mathscr{B})L + \mathbf{p}(\mathscr{B})\ell(\mathscr{B}).\end{aligned}$$

In particular, for autonomous systems $\dim(\mathscr{B}|_{[1,L]}) = \mathbf{n}(\mathscr{B})$.

Under some conditions on $w \in \mathscr{B}$ (persistency of excitation [9]), $\dim(\mathscr{B}|_{[1,L]}) = \operatorname{rank} \mathscr{H}_L(w)$ and $\mathscr{B}|_{[1,L]}$ coincides with the column space of $\mathscr{H}_L(w)$. Hence, in the context of system identification in the class $\mathscr{L}_{m,\ell}^q$ we need to adjust the rank for $L \ge \ell$ as

$$r(L) = mL + (q-m)\ell.$$

Based on this adjustment, we are able to compute initial approximations based on unstructured low-rank approximation of $\mathscr{H}_L$ for $L \ge \ell + 1$.

1) Eigenbasis initial approximation: Let

$$\widehat{U}(w_d) = \begin{bmatrix}\sigma_1 U_1 & \dots & \sigma_r U_r\end{bmatrix} \tag{12}$$

be a matrix of an approximate basis of the $\mathscr{B}|_{[1,L]}$ obtained from low-rank approximation (9). Therefore we can compute a default initial approximation from r samples of trajectories.

$$\begin{aligned}\widehat{R}_{EIG}(w_d) =& \widehat{R}_{EIG}(\widehat{U}) := \\ & \widehat{R}_{m,\ell}^{(0)}([\mathscr{H}_{l+1}(\sigma_1 U_1) \cdots \mathscr{H}_{l+1}(\sigma_r U_r)]),\end{aligned} \tag{13}$$

where $\widehat{R}_{m,\ell}^{(0)}$ is defined in (10).

2) Cadzow-based initial approximations: As it was noted in Section III-B, the Cadzow process leads to a solution $\widehat{w}_{r(L),L,\infty}^{(Cad)}(w_d)$ satisfying (11), and therefore belonging to $\mathscr{L}_{m,\ell}^q|_{[1,T]}$. Therefore we can take

$$\widehat{R}_{L,\infty}^{(Cad)}(w_d) := \widehat{R}_{m,\ell}^{(0)}(\widehat{w}_{r(L),L,\infty}^{(Cad)}(w_d))$$

as initial approximation. But we can also take approximate kernels for time series on each Cadzow iteration

$$\widehat{R}_{L,k}^{(Cad)}(w_d) := \widehat{R}_{m,\ell}^{(0)}(\widehat{w}_{r(L),L,k}^{(Cad)}(w_d)).$$

D. Special initial approximations

Here we list some special subspace-based system identification methods

1) ESPRIT initial approximation for scalar autonomous systems: The following approximation comes from signal processing ESPRIT method of estimation of a sum of damped cisoids [4], which is also connected with Kung's method of system identification [1]. Let the characteristic polynomial $R(z) = R_0 + R_1 z + \ldots + R_\ell z^\ell$ of the true trajectory w_0 have just simple complex roots.

Let $\widehat{U} \in \mathbb{R}^{L \times r}$ be a matrix of an approximate basis of $\mathscr{H}_L(w)$, and $\uparrow\widehat{U}$ and $\downarrow\widehat{U}$ denote $\widehat{U}$ with deleted last row and first row respectively. Define $\widehat{A}^{(ls)}(\widehat{U}) := (\uparrow\widehat{U})^\dagger \downarrow\widehat{U}$ be an approximate (in least squares) solution of the shift equation

$$\uparrow U \widehat{A}^{(ls)} \approx \downarrow\widehat{U}. \tag{14}$$

Then the eigenvalues $\{\widehat{\lambda_1}, \ldots, \widehat{\lambda_r}\}$ can serve as approximations of the roots of $R(z)$. Define $\widehat{R}^{(ES)} = \widehat{R}_{L,r}^{(ES)}(w) = \widehat{R}^{(ES)}(\widehat{U}(w_d))$ by

$$\widehat{R}^{(ES)}(z) := \prod_{i=1}^{r} (z - \widehat{\lambda}_i),$$

where $\widehat{U}(w_d)$ is defined in (12). For an orthogonal matrix $\widehat{U}$ the matrix $\widehat{A}^{(ls)}(\widehat{U})$ can be computed efficiently without taking a pseudoinverse [8].

2) Subspace-based approximations for input-output systems: The subspace identification methods for systems with inputs input exploit input-output partition of the variables. The most well-known are N4SID and MOESP methods (see, for example, [11, Ch. 8]). For a trajectory with fixed input/output partition $w(t) = (u(t), y(t))$, the subspace methods proceed with construction of a matrix

$$\begin{bmatrix}\mathscr{H}_L(u) \\ \mathscr{H}_L(y)\end{bmatrix},$$

which can be obtained by permuting rows of $\mathscr{H}_L(w)$ and consequent low-rank approximation. They utilise projections on row spaces of different submatrices of $\mathscr{H}_L(u)$ and $\mathscr{H}_L(y)$ in order to approximate *free responses* of the system (response of a system to zero input), which can be used to find a representation of the system (impulse response, kernel representation or input/state/output representation [11, Ch. 8]). We use a modification of the MOESP algorithm (see Appendix).

IV. Comparison of initial approximations

In this section we simulate M realisations w_k of $w_d = w_0 + \widetilde{w}$, where $\widetilde{w} \in N(0, \sigma^2)$, and compute estimates $\widehat{X}$, obtained by different methods: subspace methods and SLRA optimization started from subspace initial approximations. A natural measure of performance of an estimate is the root mean square error

$$RMSE(\widehat{X}) = \frac{\sum_{k=1}^{M} \|\widehat{X}(w_k) - X^{(0)}(w_0)\|}{M}.$$

A. Autonomous system identification

In this case $q = 1$ and $m = 0$. Consider a trajectory ($T = 400$) of a marginally stable autonomous system of order 6, given by equation

$$w_0(t+1) = 2\cos\left(\frac{2\pi t}{3}\right) + \cos\left(\frac{2\pi t}{7}\right) + \cos\left(\frac{2\pi t}{10}\right)$$

We set $M = 400$ and $L = 200$, and we denote

$$\begin{aligned} X_0 &:= \widehat{X}(\widehat{R}^{(0)}_{0,\ell}(w_d)), \\ X_k &:= \widehat{X}(\widehat{R}^{(Cad)}_{L,k}(w_d)), \\ X_{ESP} &:= \widehat{X}(\widehat{R}^{(ES)}_{L,\ell}(w_d)), \\ X_{EIG} &:= \widehat{X}(\widehat{R}_{EIG}(w_d))), \end{aligned}$$

where $\widehat{X}(\cdot)$ is defined in Remark 1. We also denote by $X_{SLRA}(X)$ the result of SLRA local optimization with initial approximation X.

In Fig. 1 we show that the estimate X_0 (default TLS approximation) deteriorates very quickly, but the other, improved, estimates have almost the same RMSE.

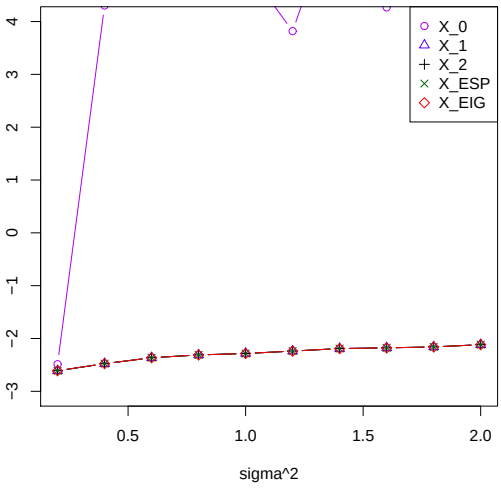

Fig. 1. $\log_{10}(RMSE)$ for the estimates depending on the noise variance

In Fig. 2 we see that the difference between improved initial approximations is very small (even on a smaller scale). The improved initial approximations lead almost always to the same solution. Therefore, we can use the computationally simplest solution. For example, we can avoid ESPRIT-type computation.

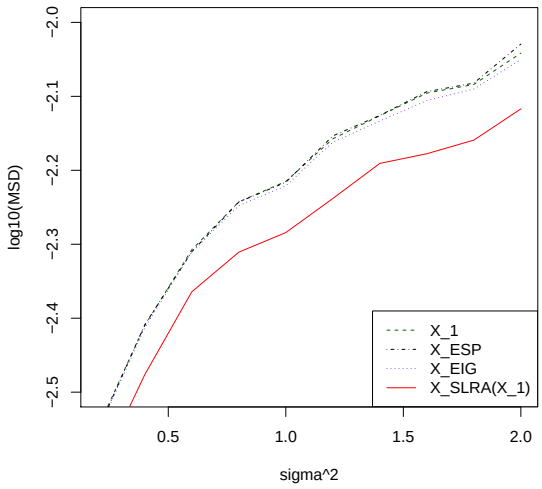

Fig. 2. $\log_{10}(RMSE)$ for the initial approximation and for the SLRA estimates started from improved initial approximation

As we see, even the first Cadzow iteration gives a good initial approximation. Hence, we also compare dependence of the quality of approximations $X_k := \widehat{X}(\widehat{R}^{(Cad)}_{L,k}(w_d))$ on the number of iteration and the window size L. In Fig. 3 we see that for $\sigma = 1$ if L is relatively small ≤ 10, then Cadzow iterations do not converge to a good initial approximation. If, however, the window length is relatively large $L \geq 20$, then one may need just one Cadzow iteration.

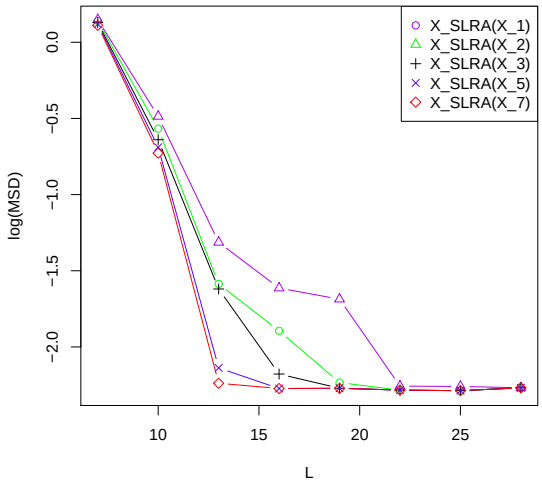

Fig. 3. Dependence of $\log_{10}(RMSE)$ on L for the SLRA started from Cadzow initial approximations, $\sigma^2 = 1$

B. System identification with inputs

We consider an example of a *SISO* (*single-input-single-output*, $q = 2$ and $m = 1$) system $w = (u_0, y_0)$ from [15]

$$y_0(t) = \frac{1\sigma^{-1} + 0.5\sigma^{-2}}{1 - 1.5\sigma^{-1} + 0.7\sigma^{-2}} u_0(t),$$

where $\widehat{X}(\widehat{w}_{1,\ell}^{(0)}) = (0.5, -0.7, 1, 1.5, 0)^{\mathrm{T}}$. The input u_0 is taken to be a particular realization of the following ARMA process $(1 + 0.7\sigma^{-1})\varepsilon(t) = (1 - 0.5\sigma^{-1})u_0(t)$, where ε is a white Gaussian noise. Note that information about input is not used in SLRA estimation (compared to [15]).

We take $T = 2000$ and $L = 300$ for Cadzow iterations, $L = 40$ for MOESP algorithm and number of realizations $M = 500$. In Fig. 4 we see that default TLS initial approximation breaks at σ^2 as initial approximation. Using Cadzow iterations does not improve the estimate. The MOESP algorithm, however, works for large noise variances and for a small window length.

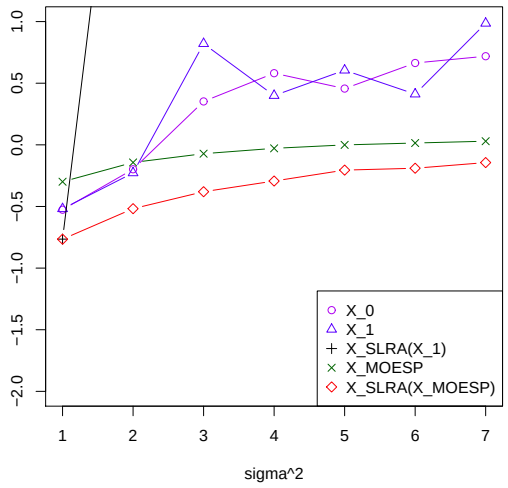

Fig. 4. Dependence of $\log_{10}(RMSE)$ on σ^2

V. CONCLUSIONS

We have proposed several initial approximations for SLRA local optimization, which are based on unstructured low-rank approximation of Hankel matrix with $L > \ell$. For scalar autonomous systems we showed that these approximations are better than the default TLS initial approximation for large noise variance. These different initial approximations yield same SLRA solutions, which allows one to take the computationally simplest initial approximation.

For a system with inputs the proposed initial approximations did not improve compared to TLS initial approximation. On the other hand, subspace methods that use information on input/output partition provide a good initial approximation. An open question remains, whether one can construct a good initial approximation that does not assume a particular input/output partition.

APPENDIX

A. MOESP method

For simulations purpose we use a slight modification of the MOESP method defined in [5], for SISO systems ($q = 2, m = 1$)

1) We first compute the QR factorization
$$[\,\mathscr{H}_L^{\mathrm{T}}(u)\mathscr{H}_L^{\mathrm{T}}(y)\,] = [\,Q_1 Q_2\,] \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix},$$
where Q_1 and Q_2 have L columns.
2) Compute the SVD of
$$(R_{22})^{\mathrm{T}} = [U_1 U_2]\Sigma[V_1 V_2]^{\mathrm{T}},$$
where $U_1 \in \mathbb{R}^{L\times\ell}$.
3) Take $(p_0, \ldots, p_{\ell-1}, p_\ell)^{\mathrm{T}} = \widehat{R}_{ES}(\widehat{A}_{ls}(U_1))$.
4) Compute impulse response $(h_0, \ldots, h_{L-1})$ from the equation
$$U_2^{\mathrm{T}} \begin{bmatrix} h_0 & 0 & \ldots & 0 \\ h_1 & h_0 & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ h_{L-1} & & & h_0 \end{bmatrix} \approx U_2^{\mathrm{T}} R_{21} (R_{11}^{-1})^{\mathrm{T}},$$
in the TLS sense.
5) Compute vector $(q_0, \ldots, q_{L-1})$ from the equation
$$\begin{bmatrix} q_\ell \\ q_{\ell-1} \\ \vdots \\ q_0 \end{bmatrix} \approx \begin{bmatrix} h_0 & 0 & \ldots & 0 \\ h_1 & h_0 & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ h_\ell & & & h_0 \end{bmatrix} \begin{bmatrix} p_\ell \\ p_{\ell-1} \\ \vdots \\ p_0 \end{bmatrix}$$
in the TLS sense.
6) Set
$$\widehat{R}_{l,L}^{(MOESP)}(w) := (q_0, p_0, \ldots, q_{\ell-1}, p_{\ell-1}, q_\ell, p_\ell)^{\mathrm{T}}$$

VI. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement no. 258581 "Structured low-rank approximation: Theory, algorithms, and applications".

We also would like to thank anonymous reviewers for their helpful comments.

REFERENCES

[1] S. Y. Kung A new identification and model reduction algorithm via singular value decomposition *Proceedings of 12th Asilomar Conference on Circuits, Systems and Computers*, 1978.

[2] J. C. Willems From time series to linear system - Part I. Finite dimensional linear time invariant systems *Automatica*, 22, 561–580, 1986.

[3] J. Cadzow Signal Enhancement - A Composite Property Mapping Algorithm *IEEE Transactions on Acoustics Speech and Signal Processing*, 36:1 49–62, 1988.

[4] A. Paulraj, R. Roy and T. Kailath Estimation of Signal Parameters via Rotational Invariance Techniques — ESPRIT. *Proceedings of 19th Asilomar Conference on Circuits, Systems and Computers*, 83–89, 1985.

[5] M. Vernaeghen and P. Dewilde Subspace model identification, Part 1: The output-error state-space model identification class of algorithms *International Journal of Control*, 56, 1187–1210, 1992.
[6] R. Pintelon, J. Schoukens *System identification: a frequency domain approach*, IEEE press, 2001.
[7] N. Golyandina, V. Nekrutkin, A. Zhigljavsky *Analysis of Time Series Structure: SSA and Related Techniques*, Chapman & Hall / CRC, 2001.
[8] R. Badeau, G. Richard, B. David Fast adaptive esprit algorithm. *2005 IEEE/SP 13th Workshop on Statistical Signal Processing*, 289–294, 17–20 July 2005.
[9] J. C. Willems, P. Rapisarda, I. Markovsky, B. L.M. De Moor A note on persistency of excitation *Systems and Control Letters*, 54 325–329, 2005.
[10] A. Kukush, I. Markovsky and S. Van Huffel Consistency of the structured total least squares estimator in a multivariate errors-in-variables model. *Journal of Statistical Planning and Inference*, 133:2 315–358, 2005.
[11] I. Markovsky, J. C. Willems, S. Van Huffel, and B. De Moor *Exact and Approximate Modeling of Linear Systems: A Behavioral Approach*. SIAM, 2006.
[12] T. Söderstrom Errors-in-variables methods in system identification *Automatica*, 43:6 939–958, 2007
[13] I. Markovsky Structured low-rank approximation and its applications. *Automatica*, 44:4 891–909, 2008.
[14] N. Golyandina On the choice of parameters in Singular Spectrum Analysis and related subspace-based methods *Statistics and Its Interface* 3, 357–368, 2010.
[15] T. Söderstrom, M. Mossberg Accuracy analysis of a covariance matching approach for identifying errors-in-variables systems *Automatica* 47, 272–282, 2011.
[16] K. Usevich and I. Markovsky Structured Low-rank Approximation as a Rational Function Minimization. *Technical report 23000*, School of Electronics and Computer Science, University of Southampton, 2011.
[17] I. Markovsky *Low Rank Approximation: Algorithms, Implementation, Applications*. Springer, 2012.