How effective is the nuclear norm heuristic in solving data approximation problems?

Ivan Markovsky *

* *School of Electronics and Computer Science, University of Southampton*
SO17 1BJ, United Kingdom, Email: im@ecs.soton.ac.uk

Abstract: The question in the title is answered empirically by solving instances of three classical problems: fitting a straight line to data, fitting a real exponent to data, and system identification in the errors-in-variables setting. The results show that the nuclear norm heuristic performs worse than alternative problem dependant methods—ordinary and total least squares, Kung's method, and subspace identification. In the line fitting and exponential fitting problems, the globally optimal solution is known analytically, so that the suboptimality of the heuristic methods is quantified.

Keywords: low-rank approximation, nuclear norm, subspace methods, system identification.

1. INTRODUCTION

With a few exceptions model reduction and system identification lead to non-convex optimization problems, for which there are no efficient global solution methods. The methods for H_2 model reduction and maximum likelihood system identification can be classified as local optimization methods and convex relaxations. Local optimization methods require an initial approximation and are in general computationally more expensive than the relaxation methods, however, the local optimization methods explicitly optimize the desired criterion, which ensures that they produce at least as good result as a relaxation method, provided the solution of the relaxation method is used as an initial approximation for the local optimization method.

A subclass of convex relaxation methods for system identification are the subspace methods, see Van Overschee and De Moor [1996]. Subspace identification emerged as a generalization of realization theory and proved to be a very effective approach. It also leads to computationally robust and efficient algorithms. Currently there are many variations of the original subspace methods (N4SID, MOESP, and CVA). Although the details of the subspace methods may differ, their common feature is that the approximation is done in two stages, the first of which is unstructured low-rank approximation of a matrix that is constructed from the given input/output trajectory.

Related to the subspace methods are Kung's method and the balanced model reduction method, which are the most effective heuristics for model reduction of linear time-invariant systems.

A recently proposed convex relaxation method is the one using the nuclear norm as a surrogate for the rank. The nuclear norm relaxation for solving rank minimization problems was proposed in Fazel et al. [2001] and was shown to be the tightest relaxation of the rank. It is a generalization of the ℓ_1-norm heuristic from sparse vector approximation problems to rank minimization problems.

The nuclear norm heuristic leads to a semidefinite optimization problem, which can be solved by existing algorithms with provable convergence properties and readily available software packages. (We use CVX, see Grant and Boyd.) Apart from theoretical justification and easy implementation in practice, formulating the problem as a semidefinite program has the additional advantage of flexibility. For example, adding regularization and affine inequality constraints in the data modeling problem still leads to semidefinite optimization problems that can be solved by the same algorithms and software as the original problem.

A disadvantage of using the nuclear norm heuristic is the fact that the number of optimization variables in the semidefinite optimization problem depends quadratically on the number of data points in the data modeling problem. This makes methods based on the nuclear norm heuristic impractical for problems with more than a few hundreds of data points. Such problems are considered "small size" data modeling problem.

Outline of the paper

The objective of this paper is to test the effectiveness of the nuclear norm heuristic as a tool for system identification and model reduction. Although, there are recent theoretical results, see, *e.g.*, Candés and Recht [2009], on exact solution of matrix completion problems by the nuclear norm heuristic, to the best of the author's knowledge there are no similar results about the effectiveness of the heuristic in system identification problems.

The nuclear norm heuristic is compared empirically with other heuristic methods on benchmark problems. The selected problems are simple: small complexity model and small number of data points. The experiments in the paper are reproducible Buckheit and Donoho [1995]. Moreover the MATLAB code that generates the results is included in the paper, so that the reader can repeat the examples by copying the code chunks from the paper and pasting them in the MATLAB command prompt, or by downloading the code from

http://eprints.soton.ac.uk/336088/

The selected benchmark problems are:

(1) line fitting by geometric distance minimization (orthogonal regression),
(2) fitting a real exponential function to data, and
(3) system identification in the errors-in-variables setting.

Problem 1 is the static equivalent of problem 3 and can be solved exactly by unstructured rank-1 approximation of the matrix of the point coordinates. Problem 2 can be viewed as a first order autonomous system identification problem. This problem also admits an exact analytic solution. Therefore in the first two cases, we are able to quantify the sub-optimality of the nuclear norm heuristic (as well as any other method). This is not possible in the third benchmark problem, where there are no methods that can efficiently compute a globally optimal point.

2. TEST EXAMPLES

2.1 Line fitting

In this section, we consider the problem of fitting a line $\mathscr{B}$, passing through the origin, to a set of points in the plain

$$\mathscr{D} = \{ d_1, \ldots, d_N \}.$$

The fitting criterion is the geometric distance from $\mathscr{D}$ to $\mathscr{B}$

$$\operatorname{dist}(\mathscr{D}, \mathscr{B}) = \sqrt{\sum_{i=1}^{N} \operatorname{dist}^2(d_i, \mathscr{B})}, \qquad \text{(dist)}$$

where

$$\operatorname{dist}(d_i, \mathscr{B}) := \min_{\widehat{d}_i \in \mathscr{B}} \| d_i - \widehat{d}_i \|_2.$$

The line fitting problem in the geometric distance sense

$$\begin{aligned} &\text{minimize} && \operatorname{dist}(\mathscr{D}, \mathscr{B}) \\ &\text{over all lines } \mathscr{B} \text{ passing through } 0 \end{aligned} \qquad \text{(LF)}$$

is equivalent to the problem of finding the nearest in the Frobenius norm $\| \cdot \|_{\mathrm{F}}$ sense rank-1 matrix $\widehat{D}$ to the matrix of the point coordinates

$$D = [d_1 \ \cdots \ d_N],$$

i.e.,

$$\begin{aligned} &\text{minimize} && \text{over } \widehat{D} \in \mathbb{R}^{q \times N} \quad \| D - \widehat{D} \|_{\mathrm{F}} \\ &\text{subject to} && \operatorname{rank}(\widehat{D}) \leq r, \end{aligned} \qquad \text{(LRA)}$$

where $q = 2$ and $r = 1$.

Note 1. (Generalization and links to other methods). For general $r < q < N$, (LRA) corresponds to fitting an r-dimensional subspace to N points in a q-dimensional space. This problem is closely related to the principal component analysis and total least squares problem Markovsky and Van Huffel [2007].

The following theorem shows that all optimal solutions of (LRA) are available analytically in term of the singular value decomposition of D.

Theorem 1. (Eckart–Young–Mirsky). Let

$$D = U \Sigma V^\top$$

be the singular value decomposition of D and partition U, $\Sigma =: \operatorname{diag}(\sigma_1, \ldots, \sigma_q)$, and V as follows:

$$\overset{\quad r \quad q-r}{U =: [U_1 \ U_2]} \ \begin{matrix} \\ q \end{matrix}, \quad \Sigma =: \begin{matrix} r \\ q-r \end{matrix} \overbrace{\begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}}^{r \quad q-r} \quad \text{and} \quad \overset{\quad r \quad q-r}{V =: [V_1 \ V_2]} \ \begin{matrix} \\ N \end{matrix},$$

Then the rank-r matrix, obtained from the truncated singular value decomposition

$$\widehat{D}^* = U_1 \Sigma_1 V_1^\top,$$

is such that

$$\| D - \widehat{D}^* \|_{\mathrm{F}} = \min_{\operatorname{rank}(\widehat{D}) \leq r} \| D - \widehat{D} \|_{\mathrm{F}} = \sqrt{\sigma_{r+1}^2 + \cdots + \sigma_q^2}.$$

The minimizer $\widehat{D}^*$ is unique if and only if $\sigma_{r+1} \neq \sigma_r$.

⟨*define* `lra` 2a⟩≡

```
function dh = lra(d, r)
[u, s, v] = svd(d);
dh = u(:, 1:r) * s(1:r, 1:r) * v(:, 1:r)';
```

Let $\widehat{D}^*$ be an optimal solution of (LRA) and let $\widehat{\mathscr{B}}^*$ be the optimal fitting model

$$\widehat{\mathscr{B}}^* = \operatorname{image}(\widehat{D}^*).$$

The rank constraint in the matrix approximation problem (LRA) corresponds to the constraint in the line fitting problem (LF) that the model $\widehat{\mathscr{B}}$ is a line passing through the origin (subspace of dimension one)

$$\dim(\widehat{\mathscr{B}}^*) = \operatorname{rank}(\widehat{D}^*).$$

We use the dimension of the model is a measure for its complexity and define the map

$$D \xrightarrow{\ \mathtt{lra}_r\ } \widehat{D}^*,$$

implemented by the function `lra`.

Let $\| D \|_*$ denotes the nuclear norm of D, *i.e.*, the sum of the singular values of D. Applying the nuclear norm heuristic to (LRA), we obtain the following convex relaxation

$$\begin{aligned} &\text{minimize} && \text{over } \widehat{D} \in \mathbb{R}^{q \times N} \quad \| \widehat{D} \|_* \\ &\text{subject to} && \| D - \widehat{D} \|_{\mathrm{F}} \leq e. \end{aligned} \qquad \text{(NNA)}$$

⟨*define* `nna` 2b⟩≡

```
function dh = nna(d, e)
cvx_begin, cvx_quiet(true);
  variables dh(size(d))
  minimize norm_nuc(dh)
  subject to
  norm(d - dh, 'fro') <= e
cvx_end
```

The parameter e in (NNA) is a user supplied upper bounds on the approximation error $\| D - \widehat{D} \|_{\mathrm{F}}$.

Let $\widehat{D}$ be the solution of (NNA). Problem (NNA) defines the map

$$D \xrightarrow{\ \mathtt{nna}_e\ } \widehat{D},$$

implemented by the function $\mathtt{nna}_e$.

> The approximation $\mathtt{nna}_e(D)$ may have rank more than r, in which case (NNA) fails to identify a valid model $\widehat{\mathscr{B}}$.

Note, however, that $\mathtt{nna}_e(D) \to 0$ as $e \to \infty$, so that for "sufficiently large" values of e, $\mathtt{nna}_e(D)$ is rank deficient (and the rank can be reduced to 0). In data modeling, the aim is to identify a model in a desired model class of bounded complexity (dimension of the model). Therefore, we are interested to characterize the set of values for the parameter e in (NNA), for which the rank constraint on $\mathtt{nna}_e(D)$ holds:

$$\mathbf{e} := \{ e \mid \operatorname{rank}\big(\mathtt{nna}_e(D)\big) \leq r \}. \qquad \textbf{(e)}$$

We hypothesise that $\mathbf{e}$ is an interval

$$\mathbf{e} = [e^*_{\mathtt{nna}}, \infty). \qquad \text{(H)}$$

The smallest value of the approximation error $\| D - \mathtt{nna}_e(D) \|_{\mathrm{F}}$, for which $\operatorname{rank}(\mathtt{nna}_e(D)) \leq r$ (*i.e.*, for which a valid model exists) characterizes the effectiveness of the nuclear norm heuristic. We define

$$\mathtt{nna}_r := \mathtt{nna}_{e^*_{\mathtt{nna}}}, \quad \text{where } e^*_{\mathtt{nna}} := \min_e \{ e \mid e \in \mathbf{e} \}.$$

A bisection algorithm for computing the limit of performance $e^*_{\mathtt{nna}}$ of the nuclear norm heuristic is given in Appendix A.

Another way to quantify the effectiveness of the nuclear norm heuristic is to compute the distance of the approximation $\mathtt{nna}_e(D)$ to the manifold of rank-r matrices

$$\varepsilon(e) = \operatorname{dist}_r\big(\mathtt{nna}_e(D)\big) := \min_{\widehat{D}} \|\mathtt{nna}_e(D) - \widehat{D}\|_{\mathrm{F}} \quad \text{subject to} \quad \operatorname{rank}(\widehat{D}) \leq r.$$

⟨*define* `dist` 3a⟩≡
```
dist = @(d, r) norm(d - lra(d, r), 'fro');
```

The function $e \mapsto \varepsilon$ presents a complexity vs accuracy trade-off in using the nuclear norm heuristic. The optimal rank-r approximation corresponds in the (ε, e) space to the point $(0, e^*_{\mathtt{lra}})$, where

$$e^*_{\mathtt{lra}} := \operatorname{dist}_r(D) = \|D - \mathtt{lra}_r(D)\|_{\mathrm{F}}.$$

The best model $\mathtt{nna}_r(D)$ identifiable by the nuclear norm heuristic corresponds to the point $(0, e^*_{\mathtt{nna}})$.

> The loss of optimality incurred by the heuristic is quantified by the difference $\Delta e_{\mathtt{nna}} = e^*_{\mathtt{nna}} - e^*_{\mathtt{lra}}$.

The following code defines a simulation example and plots the $e \mapsto \varepsilon$ function over the interval $[e^*_{\mathtt{lra}}, 1.75 e^*_{\mathtt{lra}}]$.

⟨*Test line fitting* 3b⟩≡
```
randn('seed', 0); q = 2; N = 10; r = 1;
d0 = [1; 1] * [1:N]; d = d0 + 0.1 * randn(q, N);

⟨define dist 3a⟩, e_lra = dist(d, r)
N = 20; Ea = linspace(e_lra, 1.75 * e_lra, N);
for i = 1:N
  Er(i) = dist(nna(d, Ea(i)), r);
end
figure, plot(Ea, Er, 'o', 'markersize', 8)
```

The result is shown in Figure 1. In the example,

$$e^*_{\mathtt{nna}} = 0.4603 \quad \text{and} \quad e^*_{\mathtt{lra}} = 0.3209.$$

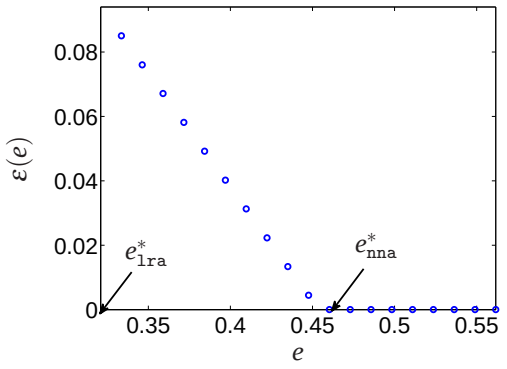

Fig. 1. Distance of $\mathtt{nna}_e(D)$ to a linear model of complexity 1 as a function of the approximation error e.

Next, we compare the loss of optimality of the nuclear norm heuristic with those of two other heuristics: line fitting by minimization of the sum of squared vertical and horizontal distances from the data points to the fitting line, *i.e.*, the classical method of solving an overdetermined linear system of equations in the least squares sense.

⟨*Test line fitting* 3b⟩+≡
```
dh_ls1 = [1; d(2, :) / d(1, :)] * d(1, :);
e_ls1  = norm(d - dh_ls1, 'fro')
dh_ls2 = [d(1, :) / d(2, :); 1] * d(2, :);
e_ls2  = norm(d - dh_ls2, 'fro')
```

The results are

$$e^*_{\mathrm{ls1}} = 0.4546 \quad \text{and} \quad e^*_{\mathrm{ls2}} = 0.4531$$

which are both slightly better than the nuclear norm heuristic.

2.2 Exponential fitting

The problem considered in this section is fitting a time series

$$y_{\mathrm{d}} := \big(y_{\mathrm{d}}(1), \ldots, y_{\mathrm{d}}(T)\big)$$

by an exponential function

$$c \exp_z := \big(cz^1, \ldots, cz^T\big)$$

in the 2-norm sense, *i.e.*,

$$\text{minimize} \quad \text{over } c \in \mathbb{R} \text{ and } z \in \mathbb{R} \quad \|y_{\mathrm{d}} - c \exp_z\|_2. \tag{EF}$$

The constraint that the sequence

$$\widehat{y} = \big(\widehat{y}(1), \ldots, \widehat{y}(T)\big)$$

is an exponential function is equivalent to the constraint that the Hankel matrix

$$\mathscr{H}_L(\widehat{y}) := \begin{bmatrix} \widehat{y}_1 & \widehat{y}_2 & \widehat{y}_3 & \cdots & \widehat{y}_{T-L+1} \\ \widehat{y}_2 & \widehat{y}_3 & \iddots & & \widehat{y}_{T-L+2} \\ \widehat{y}_3 & \iddots & & & \vdots \\ \vdots & & & & \\ \widehat{y}_L & \widehat{y}_{L+1} & \cdots & & \widehat{y}_T \end{bmatrix},$$

where $1 < L < T-1$ has rank less than or equal to 1. Therefore, the exponential fitting problem (EF) is equivalent to the Hankel structured rank-1 approximation problem

$$\begin{array}{ll} \text{minimize} & \text{over } \widehat{y} \in \mathbb{R}^T \quad \|y_{\mathrm{d}} - \widehat{y}\|_2 \\ \text{subject to} & \operatorname{rank}\big(\mathscr{H}_L(\widehat{y})\big) \leq 1. \end{array} \tag{HLRA}$$

Problem (HLRA) has analytic solution, see [De Moor, 1994, Sec. IV.C].

Lemma 1. The optimal solution of (HLRA) is

$$\widehat{y}^* = c^* \exp_{z^*}$$

where z^* is a root of the polynomial equation

$$\sum_{t=1}^{T} t y_{\mathrm{d}}(t) z^{t-1} \sum_{t=1}^{T} z^{2t} - \sum_{t=1}^{T} y_{\mathrm{d}}(t) z^t \sum_{t=1}^{T} t z^{2t-1} = 0 \tag{z^*}$$

and

$$c^* := \frac{\sum_{t=1}^{T} y_{\mathrm{d}}(t) z^{*t}}{\sum_{t=1}^{T} z^{*2t}}. \tag{c^*}$$

The proof of the lemma and an implementation of a procedure `fit_exp` for global solution of (HLRA), suggested by the lemma, are given in Appendix B.

Applying the nuclear norm heuristic to problem (HLRA), we obtain the following convex relaxation

$$\begin{array}{ll} \text{minimize} & \text{over } \widehat{y} \in \mathbb{R} \quad \|\mathscr{H}_L(\widehat{y})\|_* \\ \text{subject to} & \|y - \widehat{y}\|_2 \leq e. \end{array}$$

⟨*define* nna_exp 4a⟩≡
```
function yh = nna_exp(y, L, e)
cvx_begin, cvx_quiet(true);
  variables yh(size(y))
  minimize norm_nuc(hankel(yh(1:L), yh(L:end)))
  subject to
  norm(y - yh) <= e
cvx_end
```

As in the line fitting problem, the selection of the parameter e can be done by a bisection algorithm. As in Section 2.2, we show the complexity vs accuracy trade-off curve and quantify the loss of optimality by the difference $\Delta e_{\text{nna}} = e^*_{\text{hlra}} - e^*_{\text{nna}}$ between the optimal approximation error e^*_{hlra}, computed using the result of Lemma 1,

⟨*define* dist_exp 4b⟩≡
```
dist_exp = @(y) norm(y - fit_exp(y));
```

and the minimal error e^*_{nna}, for which the heuristic identifies a valid model.

The following code defines a simulation example and plots the trade-off curve ε over the interval $[e^*_{\text{hlra}}, 1.25e^*_{\text{hlra}}]$.

⟨*Test exponential fitting* 4c⟩≡
```
randn('seed', 0); z0 = 0.4; c0 = 1; T = 10;
t = (1:T)'; y = c0 * (z0 .^ t) + 0.1 * randn(T, 1);
⟨define dist_exp 4b⟩, e_hlra = dist_exp(y)

N = 20; Ea = linspace(e_hlra, 1.25 * e_hlra, N);
L = round(T / 2);
for i = 1:N
  Er(i) = dist_exp(nna_exp(y, L, Ea(i)));
end
ind = find(Er < 1e-6); es_nna = min(Ea(ind))
figure, plot(Ea, Er, 'o', 'markersize', 8)
```

The result is shown in Figure 2. In the example,

$$e^*_{\text{nna}} = 0.3130, \quad \text{and} \quad e^*_{\text{hlra}} = 0.2734.$$

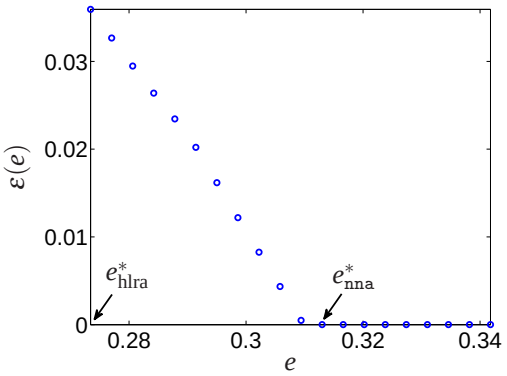

Fig. 2. Distance of $\widehat{y} = \texttt{nna}_{e^*_{\text{nna}}}(y)$ to an exponential model as a function of the approximation error $e = \|y - \widehat{y}\|$.

The performance of the nuclear norm heuristic depends on the parameter L. In the simulation example, we have fixed the value $L = \lceil T/2 \rceil$. Empirical results (see the following chunk of code and the corresponding plot in Figure 3) suggest that this is the best choice.

⟨*Test exponential fitting* 4c⟩+≡
```
Lrange = 2:(T - 1)
for L = Lrange
  Er(L) = dist_exp(nna_exp(y, L, es_nna));
end
figure,
plot(Lrange, Er(Lrange), 'o', 'markersize', 8)
```

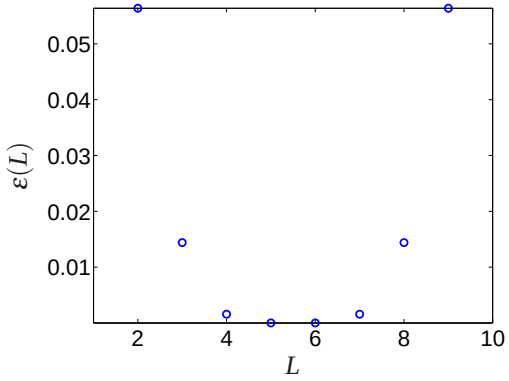

Fig. 3. Distance of $\widehat{y} = \texttt{nna}_e(y)$ to an exponential model as a function of the parameter L.

As in the line fitting problem, we compare the loss of optimality of the nuclear norm heuristic with an alternative heuristic method—Kung's method, see Kung [1978]. Kung's method is based on results from realization theory and balanced model reduction. Its core computational step is the singular value decomposition of the Hankel matrix $\mathscr{H}_L(y_{\text{d}})$, *i.e.*, unstructured low-rank approximation. The heuristic comes from the fact that the Hankel structure is not taken into account. For detailed about Kung's algorithm, we refer the reader to [Markovsky, 2012, Sect 3.1]. For completeness, an implementation kung of Kung's method is given in Appendix C.

⟨*Test exponential fitting* 4c⟩+≡
```
e_kung = norm(y - kung(y, 1, L)')
```

The obtained results is $e^*_{\text{kung}} = 0.2742$, which is much better than the result obtained by the nuclear norm heuristic.

2.3 Errors-in-variables system identification

The considered errors-in-variables identification problem is a generalization of the line fitting problem (LF) to dynamic models. The fitting criterion is the geometric distance (dist) and the model $\mathscr{B}$ is a single-input single-output linear time invariant system of order n. Let

$$w_{\text{d}} := \big(w_{\text{d}}(1), \ldots, w_{\text{d}}(T)\big), \quad \text{where } w_{\text{d}}(t) \in \mathbb{R}^2,$$

be the given trajectory of the system. The identification problem is defined as follows: given w_{d} and n,

$$\begin{aligned} &\text{minimize} && \text{dist}(w_{\text{d}}, \widehat{\mathscr{B}}) \\ &\text{subject to} && \widehat{w} \text{ is traj. of LTI system of order } n. \end{aligned} \quad \text{(SYSID)}$$

The problem is equivalent to the following block-Hankel structured low-rank approximation problem

$$\begin{aligned} &\text{minimize} && \text{over } \widehat{w} \quad \|w_{\text{d}} - \widehat{w}\|_2 \\ &\text{subject to} && \text{rank}\left(\begin{bmatrix} \mathscr{H}_L(\widehat{w}_1) \\ \mathscr{H}_L(\widehat{w}_2) \end{bmatrix}\right) \leq L + n, \end{aligned} \quad \text{(BHLRA)}$$

for $n < L < \lceil T/2 \rceil$.

⟨*define* blk_hank 4f⟩≡
```
blk_hank = @(w, L) [hankel(w(1, 1:L), w(1, L:end))
                    hankel(w(2, 1:L), w(2, L:end))];
```

This is a nonconvex optimization problem, for which there are no efficient solution methods. Using the nuclear norm heuristic, we obtain the following convex relaxation

$$\begin{array}{ll} \text{minimize} & \text{over } \widehat{w} \quad \left\| \begin{bmatrix} \mathscr{H}_L(\widehat{w}_1) \\ \mathscr{H}_L(\widehat{w}_2) \end{bmatrix} \right\|_* \\ \text{subject to} & \|w_{\mathrm{d}} - \widehat{w}\|_2 < e. \end{array}$$

⟨*define* nna_sysid 5a⟩≡
```
function wh = nna_sysid(w, L, e)
⟨define blk_hank 4f⟩
cvx_begin, cvx_quiet(true);
  variables wh(size(w))
  minimize norm_nuc(blk_hank(wh, L))
  subject to
  norm(w - wh, 'fro') <= e
cvx_end
```

A lower bound to the distance from w to a trajectory of a linear time-invariant system of order n, is given by the unstructured low rank approximation of the block Hankel matrix.

⟨*define* dist_sysid 5b⟩≡
```
⟨define blk_hank 4f⟩, ⟨define dist 3a⟩
dist_sysid = @(w, L, n) dist(blk_hank(w, L), L + n);
```

The following code defines a test example.

⟨*Test system identification* 5c⟩≡
```
randn('seed', 0); rand('seed', 0); T = 20; n = 1;
sys0 = ss(0.5, 1, 1, 1, -1); u0 = rand(T, 1);
y = lsim(sys0, u0) + 0.1 * randn(T, 1);
w = [(u0 + 0.1 * randn(T, 1))'; y'];

⟨define dist_sysid 5b⟩
N = 20; Ea = linspace(0.3, 1, N); L = 4;
for i = 1:N
  Er(i) = dist_sysid(nna_sysid(w, L, Ea(i)), L, n);
end
ind = find(Er < 1e-6); es_nna = min(Ea(ind))
figure, plot(Ea, Er, 'o', 'markersize', 8)
```

The obtained trade-off curve is shown in Figure 4. The optimal

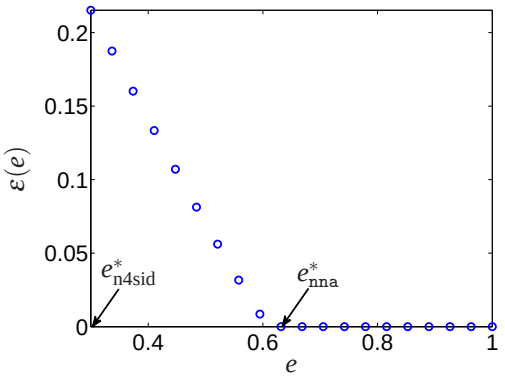

Fig. 4. Distance of $\widehat{w} = \texttt{nna}_e(w)$ to a model of order 1 as a function of the approximation error $e = \|w - \widehat{w}\|$.

model computed by the nuclear norm heuristic has corresponding approximation error $e^*_{\text{nna}} = 0.7789$. We have manually selected the value $L = 4$ as giving the best results.

⟨*Test system identification* 5c⟩+≡
```
Lrange = (n + 1):floor(T / 2);
for L = Lrange
  Er(L) = dist_sysid(nna_sysid(w, L, es_nna), L, n);
end
figure,
plot(Lrange, Er(Lrange), 'o', 'markersize', 8)
```

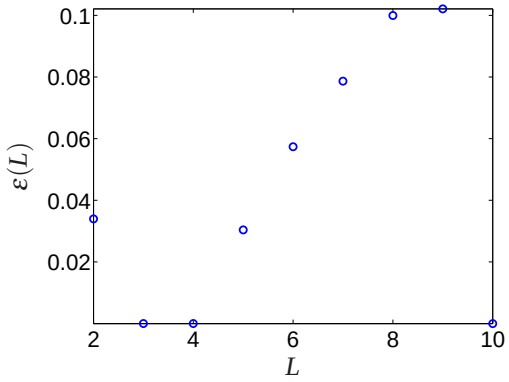

Fig. 5. Distance of $\widehat{w} = \texttt{nna}_e(w)$ to a model of order 1 as a function of the parameter L.

Next, we apply the N4SID method, implemented in function `n4sid` of the Identification Toolbox.

⟨*Test system identification* 5c⟩+≡
```
sysh = ss(n4sid(iddata(w(2,:)', w(1, :)'), ...
                n, 'nk', 0)); sysh = sysh(1, 1);
```

The distance from the data w_{d} (`w`) to the obtained model $\widehat{\mathscr{B}}$ (`sysh`) is computed by the function `misfit`, see Appendix D.

⟨*Test system identification* 5c⟩+≡
```
[e_n4sid, wh_n4sid] = misfit(w, sysh); e_n4sid
```

The approximation error achieved by the `n4sid` alternative heuristic method is $e_{\text{n4sid}} = 0.3019$. In this example, the subspace method produces a significantly better model than the nuclear norm heuristic.

3. CONCLUSIONS

The examples considered in the paper—line fitting in the geometric distance sense, optimal exponential fitting, and system identification—suggest that alternative heuristics—ordinary least squares, Kung's, and N4SID methods—are more effective in solving the original nonconvex optimization problems than the nuclear norm heuristic. Further study will focus in understanding the cause of the inferior performance of the nuclear norm heuristic and finding ways for improving it.

ACKNOWLEDGEMENTS

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 number 258581 "Structured low-rank approximation: Theory, algorithms, and applications".

REFERENCES

J. Buckheit and D. Donoho. *Wavelets and Statistics*, chapter "Wavelab and reproducible research". Springer-Verlag, Berlin, New York, 1995.

E. Candés and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9:717–772, 2009.

B. De Moor. Total least squares for affinely structured matrices and the noisy realization problem. *IEEE Trans. Signal Proc.*, 42(11):3104–3113, 1994.

M. Fazel, H. Hindi, and S. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proc. American Control Conference*, 2001.

M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming. `stanford.edu/~boyd/cvx`.

S. Kung. A new identification method and model reduction algorithm via singular value decomposition. In *Proc. 12th Asilomar Conf. Circuits, Systems, Computers*, pages 705–714, Pacific Grove, 1978.

I. Markovsky. *Low Rank Approximation: Algorithms, Implementation, Applications*. Springer, 2012.

I. Markovsky and S. Van Huffel. Overview of total least squares methods. *Signal Processing*, 87:2283–2302, 2007.

P. Van Overschee and B. De Moor. *Subspace Identification for Linear Systems: Theory, Implementation, Applications*. Kluwer, Boston, 1996.

Appendix A. BISECTION ALGORITHM FOR COMPUTING THE LIMIT OF PERFORMANCE OF NNA

Assuming that **e** is an interval (H) and observing that

$$\|D - \mathtt{lra}_r(D)\|_{\mathrm{F}} \leq e^*_{\mathtt{nna}} \leq \|D\|_{\mathrm{F}},$$

we propose a bisection algorithm on e (see Algorithm 1) for computing $e^*_{\mathtt{nna}}$.

Algorithm 1 Bisection algorithm for computing $e^*_{\mathtt{nna}}$

Input: D, r, and convergence tolerance $\varepsilon_{\Delta e}$
$e_l := \|D - \mathtt{lra}_r(D)\|_{\mathrm{F}}$ and $e_u := \|D\|_{\mathrm{F}}$
repeat
  $e := (e_l + e_u)/2$
  if $\mathrm{rank}(\mathtt{nna}_e(D)) > r$ **then**,
    $e_l := e$,
  else
    $e_u := e$.
  end if
until $\mathrm{rank}\big(\mathtt{nna}_e(D)\big) \neq r$ or $e_u - e_l > \varepsilon_{\Delta e}$
return e

⟨*bisection* 6a⟩≡
```
function e = opt_e(d, r)
el = norm(d - lra(d, r), 'fro');
eu = norm(d, 'fro');
while 1,
  e  = mean([el eu]);
  re = rank(nna(d, e), 1e-5); % numerical rank
  if re > r, el = e; else, eu = e; end
  if (re == r) && (eu - el < 1e-3), break, end
end
```

Appendix B. PROOF OF LEMMA 1 AND FUNCTION FOR GLOBAL SOLUTION OF (HLRA)

The fact that $\widehat{y}^*$ is an exponential function $c^* \exp_{z^*}$ follows from the equivalence of (HLRA) and (EF). Setting the partial derivatives of the cost function

$$f(c,z) := \sum_{t=1}^{T} \big(y(t) - cz^t\big)^2$$

of (EF) w.r.t. c and z to zero, we have the following first order optimality conditions

$$\frac{\partial f}{\partial c} = 0 \quad \Longrightarrow \quad \sum_{t=1}^{T} \big(y_{\mathrm{d}}(t) - cz^t\big) z^t = 0,$$
$$\frac{\partial f}{\partial z} = 0 \quad \Longrightarrow \quad \sum_{t=1}^{T} \big(y_{\mathrm{d}}(t) - cz^t\big) t z^t = 0.$$

Solving the first equation for c gives (c^*). The right-hand-side of the second equation is a polynomial of degree $2T$, and the resulting polynomial equation is (z^*).

⟨*Analytic solution of the exponential fitting problem* 6b⟩≡
```
function [yh, ch, zh] = fit_exp(y)
t  = (1:length(y))';
p1(t) = t .* y(t); p2(2 * t + 1) = 1;
p3(t + 1) = y(t); p4(2 * t) = t;
r = roots(conv(p1, p2) - conv(p3, p4));
r(r == 0) = []; z = 1 ./ r(imag(r) == 0);
for i = 1:length(z)
  c(i) = (z(i) .^ t) \ y;
  Yh(:, i) = c(i) * (z(i) .^ t);
  f(i) = norm(y - Yh(:, i)) ^ 2;
end
[f_min, ind] = min(f);
zh = z(ind); ch = c(ind); yh = Yh(:, ind);
```

Appendix C. IMPLEMENTATION OF KUNG'S METHOD

⟨*Kung method* 6c⟩≡
```
function yh = kung(y, n, L)
[U, S, V] = svd(hankel(y(1:L), y(L:end)));
O = S(1:n, 1:n) * U(:, 1:n); C = V(:, 1:n)';
c = O(1, :); b = C(:, 1);
a = O(1:end - 1, :) \ O(2:end, :);
for t = 1:length(y)
  yh(t) = c * (a ^ (t - 1)) * b;
end
```

Appendix D. DISTANCE COMPUTATION IN THE DYNAMIC CASE

The problem of computing the distance, also called misfit, from a time series to a linear time-invariant model is a convex quadratic optimization problem. The solution is therefore available in closed form. The linear time-invariant structure of the system, however, allows efficient $O(T)$ computation, *e.g.*, the Kalman smoother computes the misfit in $O(T)$ flops, using on a state space representation of the system.

In [Markovsky, 2012, Section 3.2], a method based on, what is called an image representation of the system is presented. The following implementation does not exploit the structure in the problem and the algorithm has computational cost $O(T^3)$.

⟨*define* `misfit` 6d⟩≡
```
function [M, wh] = misfit(w, sys)
⟨transfer function to image representation 6e⟩
T  = size(w, 2); TP = blktoep(P, T);
wh = reshape(TP * (TP \ w(:)), 2, T);
M  = norm(w - wh, 'fro');
```

The function `blktoep` (not shown) constructs a block Toeplitz matrix and the conversion from transfer function to image representation is done as follows:

⟨*transfer function to image representation* 6e⟩≡
```
[q, p] = tfdata(tf(sys), 'v');
P = zeros(2, length(p));
P(1, :) = fliplr(p);
P(2, :) = fliplr([q zeros(length(p) - length(q))]);
```