INFORMATICA, 2011, Vol. 22, No. 4, 489–505

Analysis of Structured Low Rank Approximation as an Optimization Problem

Jonathan GILLARD, Anatoly ZHIGLJAVSKY
Cardiff School of Mathematics, Cardiff University
Senghenydd Road, Cardiff, UK
e-mail: {gillardjw, zhigljavskyaa}@cardiff.ac.uk

Received: June 2011; accepted: September 2011

Abstract. In this paper, we consider the so-called structured low rank approximation (SLRA) problem as a problem of optimization on the set of either matrices or vectors. Briefly, SLRA is defined as follows. Given an initial matrix with a certain structure (for example, Hankel), the aim is to find a matrix of specified lower rank that approximates this initial matrix, whilst maintaining the initial structure. We demonstrate that the optimization problem arising is typically very difficult; in particular, the objective function is multiextremal even in simple cases. We also look at different methods of solving the SLRA problem. We show that some traditional methods do not even converge to a locally optimal matrix.

Keywords: structured low rank approximation, Hankel matrix, optimization.

1. Introduction

1.1. *Statement of the Problem*

Let L and K be two given integers with $L \leqslant K$ and $\mathbb{R}^{L\times K}$ be the set of all real-valued $L \times K$ matrices. Let $\mathcal{M}_r \subset \mathbb{R}^{L\times K}$ be the subset of $\mathbb{R}^{L\times K}$ containing matrices with rank $\leqslant r$, and $\mathcal{H} \subset \mathbb{R}^{L\times K}$ be the subset of $\mathbb{R}^{L\times K}$ containing matrices of a specified structure. Define $\mathcal{A} = \mathcal{M}_r \cap \mathcal{H}$.

Assume we are given a matrix $\mathbf{X}_0 \in \mathbb{R}^{L\times K}$. The problem of structured low rank approximation (SLRA) is:

$$F(\mathbf{X}) \to \min_{\mathbf{X}\in\mathcal{A}}, \tag{1}$$

where $F(\mathbf{X}) = \rho(\mathbf{X}, \mathbf{X}_0)$ is a distance function on $\mathbb{R}^{L\times K} \times \mathbb{R}^{L\times K}$.

We only consider the case where $\mathcal{H}$ is a set of Hankel matrices (a matrix $\mathbf{X} = (x_{ij})$ is Hankel if $x_{ij} = \text{const}$ for all pairs (i, j) such that $i + j = \text{const}$). In this case, the SLRA problem (1) is called Hankel SLRA. This is the most common instance of the SLRA problem. Other structures that appear in relation to the problem (1) include, amongst others, Toeplitz, circulant, and Hankel-block-Hankel (Markovsky, 2008).

We will show that (1) is a difficult optimization problem. We will also look at different methods of solving (1). One of the main difficulties in optimising the objective function $F(\cdot)$ in (1) is that although $\mathcal{H}$ is a convex space, $\mathcal{M}_r$ is non-convex. In particular, we demonstrate that the SLRA problem (1) is generally a global, rather than a local optimization problem. Moreover, we show that the multi-extremality of the objective function is very high and the Lipschitz constant is typically very large which makes many deterministic methods of global optimization almost unsuitable. This opens huge possibilities for the stochastic methods of global optimization. Note that the seminal book by Mockus (1989) was at the heart of the field at the early stages of the development of the field of stochastic global optimization. Many different and somehow more mature methods have been developed since Clausen and Žilinskas (2002), and Calvin and Žilinskas (2000). For a comprehensive exposition of the field of stochastic global optimization see Zhigljavsky (1991) or Zhigljavsky and Žilinskas (2008).

Many problems in system identification, signal processing and time series can be reduced to the Hankel SLRA problem (1) Markovsky *et al.* (2006). Indeed, a vector of observations (called time series in statistical literature) can be mapped onto a matrix created by a series of lagged vectors. This matrix is Hankel and is in one-to-one correspondence with the original vector of observations. See Golyandina (2010) for a recent review of the corresponding techniques.

The relation between low-rank Hankel matrices and vectors governed by linear recurrent formulae has long been known; see for example (Gantmacher, 1959; Pollock, 1999). One can reformulate the original problem (1) as a problem of optimizing coefficients of linear recurrent formulae. However, in this paper we adopt a more traditional approach which is based on the representation (4) below.

Abatzoglou *et al.* (1991) are often cited to be the first to formulate a version of the SLRA problem. Since then, a number of papers discussing and offering solutions to the SLRA problem have appeared. See Markovsky (2010) for a historical overview and bibliography of important results and papers. Further details are provided in another paper by Markovsky (2008). A general perspective on the topic is also offered by Chu *et al.* (2003).

1.2. *Types of Algorithms*

As $\mathcal{A}$ is a closed set, the solution to the original problem (1) always exists. However, the solution is not necessarily unique. Set

$$\mathfrak{X}_* = \left\{ \mathbf{X}_* = \arg\min_{\mathbf{X}\in\mathcal{A}} \rho(\mathbf{X}, \mathbf{X}_0) \right\} \quad \text{and} \quad F_* = F(\mathbf{X}_*) = \min_{\mathbf{X}} F(\mathbf{X}).$$

A result of the application of an optimization algorithm to (1) is a sequence of matrices $\mathbf{X}_1, \mathbf{X}_2, \ldots$ such that some of the matrices $\mathbf{X}_n$ for large n can be considered as approximations to $\mathbf{X}_*$, a solution of (1). The optimization algorithm converges if

$$\liminf_{n\to\infty} \rho(\mathbf{X}_0, \mathbf{X}_n) = F_* \quad \Longleftrightarrow \quad \liminf_{n\to\infty} \rho(\mathbf{X}_n, \mathfrak{X}_*) = 0.$$

We propose to distinguish between the following two types of algorithms for solving problem (1).

1. Any Hankel matrix $\mathbf{X}$ of size $L \times K$ is in one-to-one correspondence with a vector of size $N = L + K - 1$. Therefore, if an algorithm generates a sequence of matrices $\mathbf{X}_n \in \mathcal{H}$ then it can also be considered as an algorithm which generates a sequence of vectors $(f_1^{(n)}, \ldots, f_N^{(n)}) \in \mathbb{R}^N$. Any solution of problem (1) can also be regarded as an N-dimensional vector. Note however that to make a transition $\mathbf{X}_n \to \mathbf{X}_{n+1}$ we may use additional matrices that do not belong to the space $\mathcal{H}$. These algorithms, where $\mathbf{X}_n \in \mathcal{H}$ for all n, can alternatively be considered as algorithms of optimization of N-dimensional functions.
2. The matrices $\mathbf{X}_n$ do not belong to the space $\mathcal{H}$ so that we approach a solution $\mathbf{X}_* \in \mathfrak{X}_*$ with a sequence of matrices $\mathbf{X}_n \notin \mathcal{H}$. In this case, the algorithms cannot be considered as optimization algorithms in $\mathbb{R}^N$ but rather as interior point algorithms where it is natural to give penalties for violating the low-rank and/or Hankel constraints.

One could think about other types of algorithms; for example, the algorithms which force the matrices $\mathbf{X}_n$ to belong to $\mathcal{M}_r$ for all $n \geqslant 1$. Such algorithms do not seem to be practical, however.

A few algorithms of solving the SLRA problem are known, see below. One of the simplest and most popular of them is the so-called Cadzow iterations algorithm, see Section 3.4. Despite Cadzow's initial claim that it is a globally convergent algorithm, this algorithm is not even locally convergent. Some other algorithms known in the literature (like HTLS, again see Section 3.4) are typically locally convergent but as shown below are not necessarily globally convergent. In this paper, we investigate the reasons for this. In particular, we show that typically the objective function $F(\cdot)$ has a large number of local minima.

1.3. *Choice of a Distance Function*

Most commonly, the distance function $\rho(\cdot, \cdot)$ is defined by the matrix Frobenius norm; that is, $\rho(\mathbf{X}, \mathbf{X}_0) = \|\mathbf{X} - \mathbf{X}_0\|_F$. In this case, the Hankel SLRA problem is equivalent to the so-called Hankel total least squares (HTLS) problem; for further information see Golyandina (2010). Consider a Hankel matrix $\mathbf{X} \in \mathbb{R}^{L \times K}$ and its vector representation $(f_1, \ldots, f_N) \in \mathbb{R}^N$ with $N = L+K-1$. Then the (i, j)th element of $\mathbf{X}$ is $x_{ij} = f_{i+j-1}$ and the squared Frobenuis norm of $\mathbf{X}$ is

$$\|\mathbf{X}\|_F^2 = \sum_{i,j=1}^{L,K} x_{ij}^2 = \sum_{m=1}^{N} w_m f_m^2, \tag{2}$$

where $\{w_1, \ldots, w_N\}$ is the sequence of weights defined by

$$w_m = \begin{cases} m, & \text{if } m \leqslant L, \\ L, & \text{if } L \leqslant m \leqslant K, \\ N - m + 1, & \text{if } K \leqslant m \leqslant N. \end{cases}$$

Another common choice of the distance function arises from the simple L_2 vector norm of the associated vector $(f_1, \ldots, f_N)$:

$$\|\mathbf{X}\|^2 = \sum_{m=1}^{N} f_m^2. \tag{3}$$

In this case, the weights attributed to each element of the vector $(f_1, \ldots, f_N)$ are equal to 1. Below we use both these norms, (2) and (3).

2. Study of the Properties of the Optimization Problem (1)

2.1. *A Parametrization of the Solution of the Problem* (1)

An alternative way of formulating the problem of Hankel SLRA (1) is by associating matrices $\mathbf{X} \in \mathcal{A}$ with vectors $(f_1, \ldots, f_N)$ which can be represented as sums of damped sinusoids:

$$f_j = \sum_{i=1}^{q} a_i \exp(d_i j) \sin(2\pi\omega_i j + \phi_i), \quad j = 1, \ldots, N; \tag{4}$$

see, for example, Lemmerling and Van Huffel (2001), and Van Huffel (1993). Here $q = r/2$ (assuming r is even) and $a = (a_1, \ldots, a_q)$, $d = (d_1, \ldots, d_q)$, $\omega = (\omega_1, \ldots, \omega_q)$ and $\phi = (\phi_1, \ldots, \phi_q)$ are sets of parameters. This formulation of the problem offers a convenient way of visualizing rank deficient Hankel matrices.

Assume that $a^{(0)} = (a_1^{(0)}, \ldots, a_q^{(0)})$, $d^{(0)} = (d_1^{(0)}, \ldots, d_q^{(0)})$, $\omega^{(0)} = (\omega_1^{(0)}, \ldots, \omega_q^{(0)})$ and $\phi^{(0)} = (\phi_1^{(0)}, \ldots, \phi_q^{(0)})$ are true values of the parameters. If the observations are noise-free, then the vector $(f_1, \ldots, f_N)$ with components

$$f_j = \sum_{i=1}^{q} a_i^{(0)} \exp\left(d_i^{(0)} j\right) \sin\left(2\pi\omega_i^{(0)} j + \phi_i^{(0)}\right), \quad j = 1, \ldots, N, \tag{5}$$

is considered as given and the objective function in the Hankel SLRA problem, with the distance function given by the norm (3), becomes

$$F(a, d, \omega, \phi) = \sum_{j=1}^{N} \varepsilon^2(j, a, d, \omega, \phi), \tag{6}$$

where

$$\varepsilon(j, a, d, \omega, \phi) = f_j - \sum_{i=1}^{q} a_i \exp(d_i j) \sin(2\pi\omega_i j + \phi_i). \tag{7}$$

If q is known then $F_* = 0$. With noise, the value f_j are not observed, but rather $f_j + n_j$, where $\{n_j,\ j = 1, \ldots, N\}$ is the series of noise terms. In this case, the objective function is also defined by (6) and (7) but typically $F_* > 0$.

We are aware of only one paper, which is Lemmerling and Van Huffel (2001), that contains a discussion about the behaviour of the objective function (6). In Lemmerling and Van Huffel (2001), the fact that the objective function F is multiextremal has been observed; the function F was decomposed into three different components and it was numerically demonstrated that the part of the objective function with the observation noise removed dominates the shape of the objective function. In this section we are not going to pursue this issue but instead provide some simple examples in order to visualize the complexity of the problem of minimization of the objective function (6).

The function F defined in (6) has the most interesting behaviour as a function of ω. For simplicity we assume that the true values of the parameters a, d and ϕ are known and are $a_i^{(0)} = 1,\ d_i^{(0)} = 0$, and $p_i^{(0)} = 0$ for $i = 1, \ldots, q$. Therefore, the objective function F we consider in the examples below is

$$F(\omega) = \sum_{j=1}^{N} \left(f_j - \sum_{i=1}^{q} \sin(2\pi\omega_i j) \right)^2, \tag{8}$$

where

$$f_j = \sum_{i=1}^{q} \sin\left(2\pi\omega_i^{(0)} j\right), \quad j = 1, \ldots, N, \tag{9}$$

with q and $\omega_1^{(0)}, \ldots, \omega_q^{(0)}$ given. In the examples with noisy observations we shall use the objective function (8) with $f_j + n_j$ substituted for f_j, where n_j represents a noise-term.

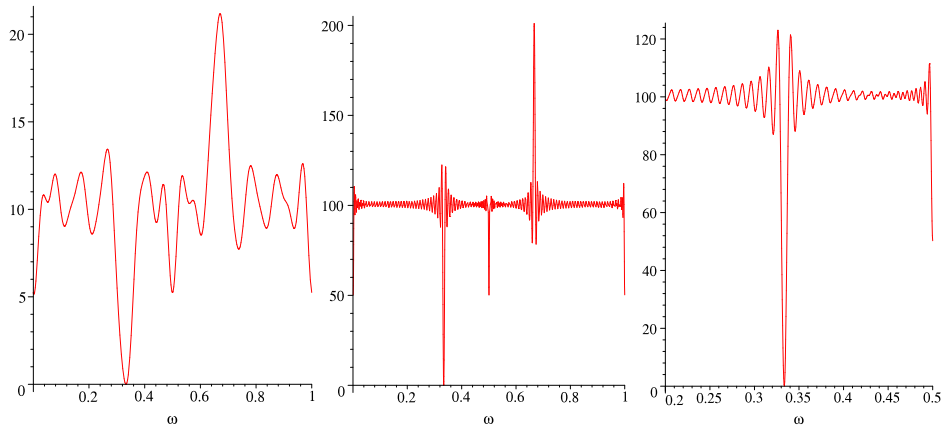
2.2. *Examples with* $q = 1$

EXAMPLE 1. $q = 1$, $N = 10$ and $N = 100$, $\omega_1^{(0)} = 1/3$; noise-free observations and noisy observations.

The objective function becomes $F(\omega) = \sum_{j=1}^{N} \left(f_j - \sin(2\pi\omega j)\right)^2$, where $\omega = \omega_1$. The function $F(\omega)$ is periodic with period 1 and the minimal value $F_* = 0$ attained at the points $\omega_1^* = \omega_1^{(0)} + k$ ($k = 0, \pm 1, \pm 2, \ldots$). The feasible domain for ω can therefore be chosen as $0, 1)$; in this interval, the function F has one global minimizer at $\omega_1^* = \omega_1^{(0)}$ and many local minimizers.

The behaviour of the function $F(\omega)$ is illustrated in Fig. 1 for $N = 10$ and $N = 100$. We can easily see that the objective function $F(\omega)$ is multiextremal and very irregular. For $N = 10$, the Lipschitz constant of F is approximately 327.86. For $N = 100$, the Lipschitz constant F is approximately 6195.88.

The global minimum, however, is visible and reasonably well-separated (in terms of the objective function values) from the multitude of local minima. Since the global



(a) $N = 10$ (b) $N = 100$ (c) $N = 100$, narrower scale

Fig. 1. Function $F(\omega)$ in Example 1.

minimizer has a very narrow domain of attraction, particular care should be taken in devising global optimization algorithms that would be able to find this global minimizer. If we know that the observations are noise-free and hence $F_* = 0$, the creation of such algorithms is not a serious challenge though.

Increasing N leads to more erratic cost function; as N increases, the number of local minima also increases. For the parameter settings in this example, the number of local minimizers increases linearly in N. In this and similar examples, the number of minimizers of $F(\omega)$, for $\omega \in [0, 1)$, is in between N and $2N$ for all $N > 3$.

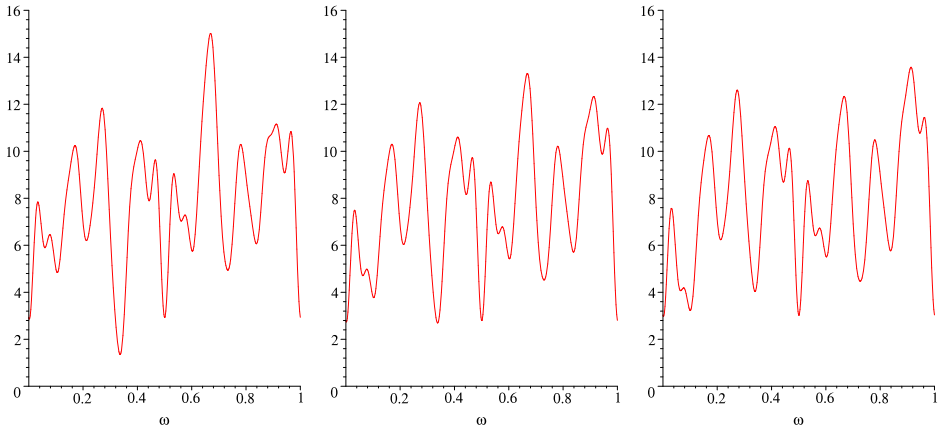
Now assume that the values f_j are not observed, but rather $f_j + n_j$, where $\{n_j, j = 1, \ldots, N\}$ is the series of uncorrelated noise terms such that n_j are normally distributed with mean 0 and variance σ^2. Figure 2 contains plots of $F(\omega)$ for particular realizations of noise for varying values of σ^2. A plot of one of the observed series (with $N = 10$) $f_j + n_j$ is included in Fig. 3.

2.3. *Examples with* $q = 2$

EXAMPLE 2. $q = 2$, $N = 10$, $\omega_1^{(0)} = 0.3$, $\omega_2^{(0)} = 0.32$; noise-free observations. The objective function is

$$F(\omega) = F(\omega_1, \omega_2) = \sum_{j=1}^{N} \big(f_j - \sin(2\pi\omega_1 j) - \sin(2\pi\omega_2 j)\big)^2,$$
$$\omega \in [0, 1) \times [0, 1). \tag{10}$$

Figure 4 contains various plots of the objective function F. In this example, two 'close' frequencies are selected. The objective function $F(\omega)$ is highly multiextremal. The global minimum is again clearly visible and well-separated (in terms of objective



(a) $\sigma^2 = 0.1, \omega_1^* = 0.3362$, $F(\omega_1^*) = 1.35, F(\omega_1^0) = 1.41$ (b) $\sigma^2 = 0.2, \omega_1^* = 0.3377$, $F(\omega_1^*) = 2.70, F(\omega_1^0) = 2.83$ (c) $\sigma^2 = 0.3, \omega_1^* = 0.5619$, $F(\omega_1^*) = 3.91, F(\omega_1^0) = 4.24$

Fig. 2. Function $F(\omega)$ in Example 1 with noisy observations, $N = 10$.

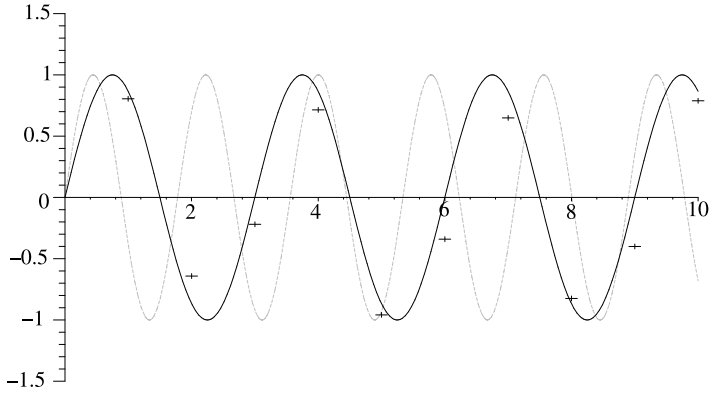
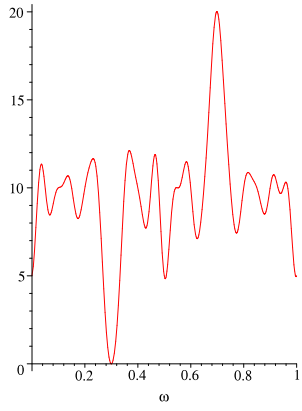


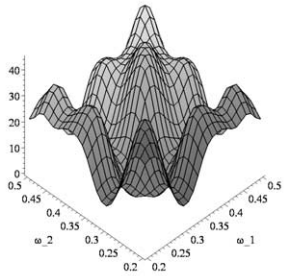
Fig. 3. Plot of $\sin(2\pi\omega_1^{(0)} j)$ (black), and $\sin(2\pi\omega_1^* j)$ (gray), where $\omega_1^{(0)} = 0.3$ and $\omega_1^* \cong 0.5619$ is the global minimizer of F for $\sigma^2 = 0.3$ and the realization related to Fig.2.

function values) from the multitude of local minima, but the effect of having two nearly equal frequencies can be seen.

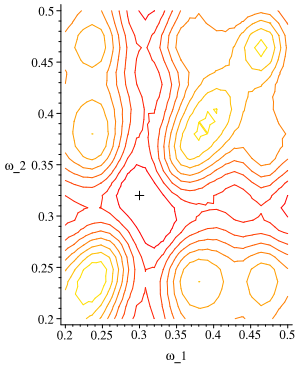
EXAMPLE 3. $q = 2, N = 10, \omega_1^{(0)} = 0.3, \omega_2^{(0)} = 0.45$; noise-free observations. The objective function and the feasible domain are as defined in (10). Figure 5 contains various plots of the objective function F. Figure 6 contains cross-sections of F as functions of ω_1 for different values of ω_2. The global minimum $F_* = 0$ of F at $\omega_1 = 0.3, \omega_2 = 0.45$ is clearly visible. Figure 6 illustrates that there is a rapid change of the objective function behaviour as ω_1, ω_2 move away from the optimal values.



(a) Plot of $F(\omega_1, 0.32)$

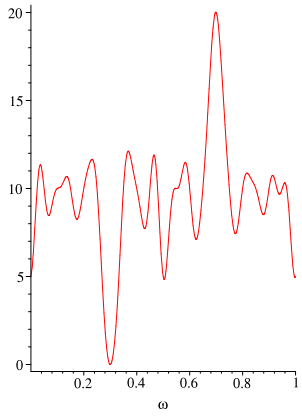


(b) Plot of $F(\omega_1, \omega_2)$

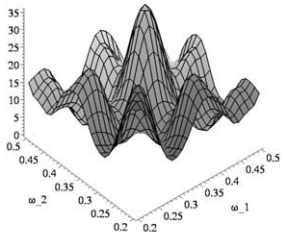


(c) Contourplot of $F(\omega_1, \omega_2)$ with the global minimizer $(\omega_1^{(0)}, \omega_2^{(0)}) = (0.3, 0.32)$ marked (+)

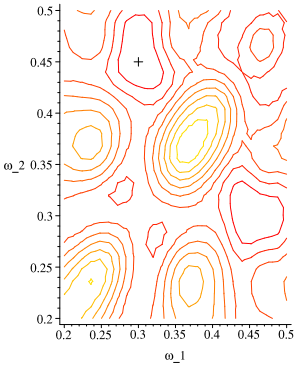
Fig. 4. Example 2. Function $F(\omega)$.



(a) Plot of $F(\omega_1, 0.45)$



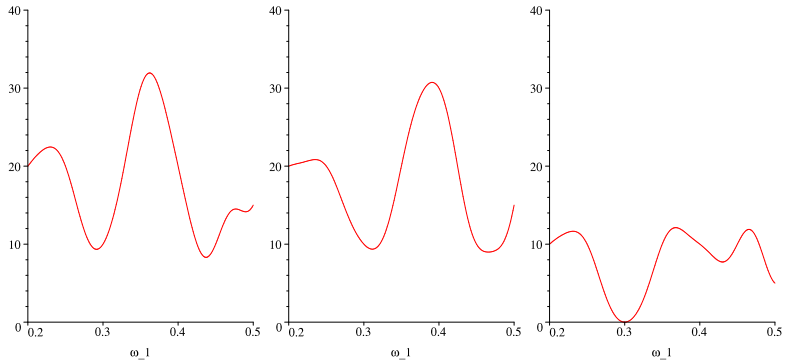
(b) Plot of $F(\omega_1, \omega_2)$



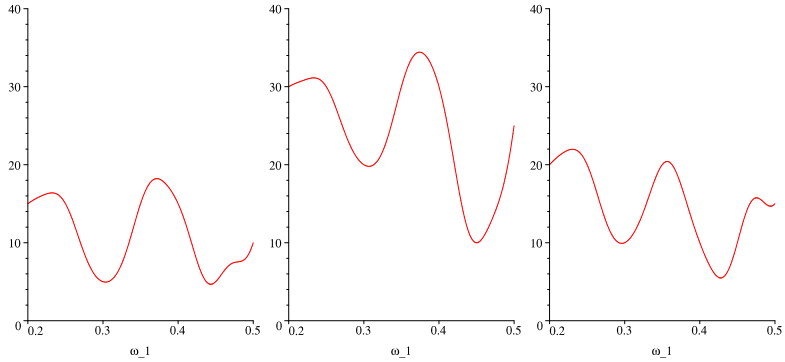
(c) Contourplot of $F(\omega_1, \omega_2)$ with the global minimizer $(\omega_1^{(0)}, \omega_2^{(0)}) = (0.3, 0.45)$ marked (+)

Fig. 5. Example 3. Function $F(\omega)$.

EXAMPLE 4. $q = 2$, $N = 10$, $\omega^{(0)} = (0.3, 0.45)$; noisy observations. In this example, we assume that f_j are not observed, but rather $f_j + n_j$ where $\{n_j, j = 1, \ldots, N\}$ is the series of uncorrelated normally distributed noise terms with variance σ^2. Figure 7 contains contourplots of $F(\omega_1, \omega_2)$ for varying values of σ^2. Plot of one of the realizations of the observed series $f_j + n_j$ is included in Fig. 8. Values of F at $\omega^{(0)}$ and the global minimizer ω^* are provided.



(a) $\omega_2 = 0.35$ (b) $\omega_2 = 0.4$ (c) $\omega_2 = 0.45$



(d) $\omega_2 = 0.5$ (e) $\omega_2 = 0.55$ (f) $\omega_2 = 0.6$

Fig. 6. Cross-sections $\tilde{F}(\omega_1) = F(\omega_1, \omega_2)$ for different values of ω_2 and a narrow scale for ω_1.

3. Methods and Their Properties

3.1. *Cadzow Iterations*

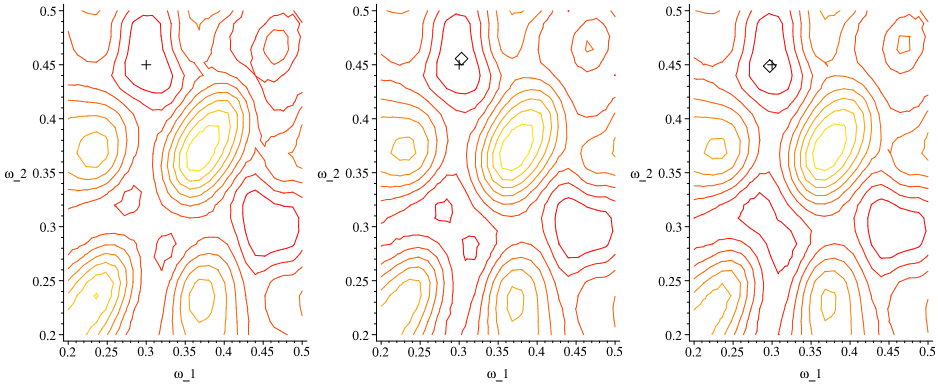
A simple algorithm where $\mathbf{X}_n \in \mathcal{H}$ for all n is provided by the use of alternating projections where to get $\mathbf{X}_{n+1}$ we project each $\mathbf{X}_n$ to the space $\mathcal{M}_r$ and then back to $\mathcal{H}$. If $\rho(\cdot, \cdot)$ is the distance induced by the Frobenius norm, then the projection to the space $\mathcal{M}_r$ is obtained by the singular value decomposition (SVD) of $\mathbf{X}_n$. The resulting algorithm is called Cadzow iterations.

Cadzow's algorithm decomposes the SLRA into two smaller problems; namely that of

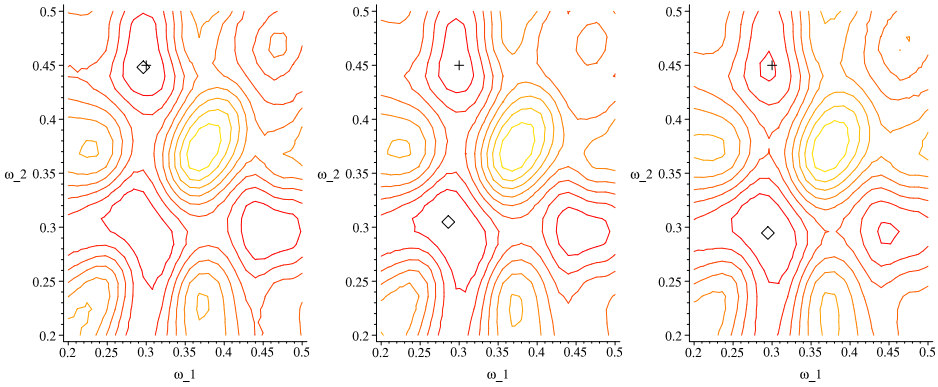
(1) finding a low-rank matrix approximation of $\mathbf{X}_n$ (using the SVD);

(2) finding a structured Hankel matrix closest to the low-rank approximation of $\mathbf{X}_0$.

One Cadzow iteration corresponds to a technique known as singular spectrum analysis (SSA); for further details regarding the link between Cadzow iterations and SSA, see Gillard (2010), and Golyandina *et al.* (2010).



(a) $\sigma^2 = 0, F(\omega^{(0)}) = 0$ (b) $\sigma^2 = 0.3, \omega^* = (0.3030, 0.4556)$, $F(\omega^*) = 0.68$, $F(\omega^{(0)}) = 1.15$ (c) $\sigma^2 = 0.5, \omega^* = (0.2970, 0.4487)$, $F(\omega^*) = 6.98$, $F(\omega^{(0)}) = 7.72$



(d) $\sigma^2 = 1, \omega^* = (0.2963, 0.4482)$, $F(\omega^*) = 13.98$, $F(\omega^{(0)}) = 23.12$ (e) $\sigma^2 = 2, \omega^* = (0.2860, 0.3049)$, $F(\omega^*) = 26.27$, $F(\omega^{(0)}) = 38.12$ (f) $\sigma^2 = 4, \omega^* = (0.2947, 0.2947)$, $F(\omega^*) = 49.59$, $F(\omega^{(0)}) = 74.27$

Fig. 7. Contour plots of $F(\omega_1, \omega_2)$ with $\omega^{(0)} = (0.3, 0.45)$ (marked $+$) and the global minimizer ω^* (marked $\diamond$), for different values of σ^2. Values of F at $\omega^{(0)}$ and ω^* are provided.

The paper by Gillard (2010) contains some examples comparing SSA with Cadzow iterations. In the simulation study within this paper, it was shown that repeated iterations of Cadzow's algorithm may result in an increased root mean square error from the true signal. This was particularly the case if the number of terms selected in the truncated SVD was smaller than the rank of the true signal.

Cadzow's algorithm is a simple heuristic method, which is suboptimal in terms of the Frobenius optimality criterion (Gillard, 2010). The example used by De Moor (1994) demonstrates the sub-optimality of Cadzow's algorithm; this example is simplistic however.

3.2. *HSVD and HTLS*

Both the HSVD and HTLS algorithms are described in a paper by Van Huffel (1993). Both of these methods parameterise the space of rank r Hankel matrices by writing the

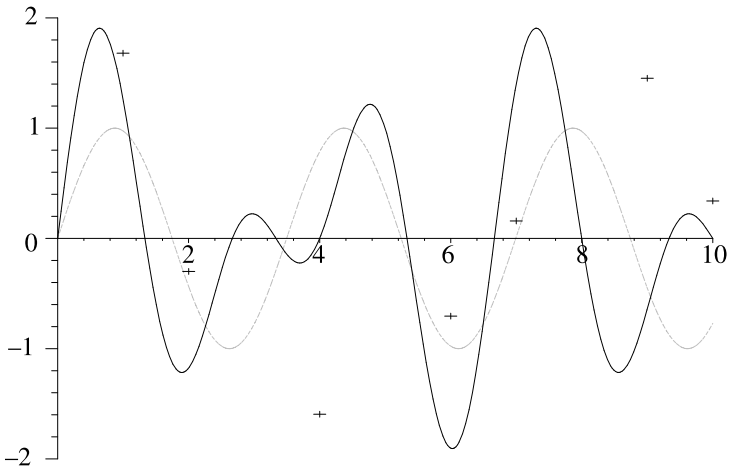


Fig. 8. Plot of $\sin(2\pi\omega_1^{(0)}j) + \sin(2\pi\omega_2^{(0)}j)$ (black), and $\sin(2\pi\omega_1^*j) + \sin(2\pi\omega_2^*j)$ (grey) where $(\omega_1^*, \omega_2^*) \cong (0.2860, 0.3049)$ is the global minimizer of F for $\sigma^2 = 2$.

signal $(f_1, \ldots, f_n)$ in the form as given in (5). Both algorithms are briefly described below.

The SVD of $\mathbf{X}_0$ gives

$$\mathbf{X}_0 = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^T.$$

Set the smallest singular values $\sqrt{\lambda_{r+1}}, \ldots, \sqrt{\lambda_d}$ to 0 resulting in the closest rank r approximation (with respect to Frobenius norm) to $\mathbf{X}_0$, denoted $\mathbf{X}$:

$$\mathbf{X} = \sum_{i=1}^{r} \sqrt{\lambda_i} U_i V_i^T$$

Let $U = [U_1 : \ldots : U_r]$, and let $\overline{U}$, $\underline{U}$ denote U with its first and last row omitted respectively. Both algorithms HSVD and HTLS involve finding the solution Y of

$$\underline{U} Y = \overline{U}. \tag{11}$$

If (11) is solved by the least squares, then the algorithm is known as HSVD. If (11) is solved by the total least squares, the algorithm is known as HTLS. Once Y is estimated, it is possible to find estimates of the vectors a, d, m, p of (4). The least squares and total least squares algorithms are described briefly in the next section.

3.3. *Least Squares and Total Least Squares Algorithms and Their Relation to SLRA*

Let $\mathbf{A} \in \mathbb{R}^{L \times M}$ be an input data matrix, and $\mathbf{B} \in \mathbb{R}^{L \times r}$ be an output observation matrix. Finding an approximation $\theta \in \mathbb{R}^{M \times r}$ such that $\mathbf{A}\theta = \mathbf{B}$ in an overdetermined system

($L > M$) is a common problem that has many applications in a variety of scientific disciplines. Common methods of solution involve perturbing $\mathbf{B}$, or both $\mathbf{A}$ and $\mathbf{B}$ as little as possible so that the new system of equations has an exact solution. There are two options:

$$\hat{\theta} = \arg\min_{\theta, \Delta\mathbf{B}} \|\Delta\mathbf{B}\| \quad \text{such that} \quad \mathbf{A}\theta = \mathbf{B} + \Delta\mathbf{B}, \tag{12}$$

or

$$\hat{\theta} = \arg\min_{\theta, \Delta\mathbf{A}\Delta\mathbf{B}} \| [\Delta\mathbf{A}\Delta\mathbf{B}] \| \quad \text{such that} \quad (\mathbf{A} + \Delta\mathbf{A})\theta = \mathbf{B} + \Delta\mathbf{B}. \tag{13}$$

If the Frobenius norm is chosen, then these methods of solution, (12) and (13), are known as least squares and total least squares, respectively.

If A is of full rank, then the least squares estimator for θ is given by $\hat{\theta}_{LS} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{B}$. There are many modifications based on this which under some circumstances, may be more robust. For example, ridge regression which imposes a constraint on the size of $\hat{\theta}$, the lasso, and the elastic net; see Hastie *et al.* (1998).

It is possible to rewrite (13) as a matrix low rank approximation problem. Let $\mathbf{X} = [\mathbf{AB}] \in \mathbb{R}^{L\times K}$, then for suitable r and a matrix norm $\|\cdot\|$ find $\hat{\mathbf{X}}$ such that

$$\|\mathbf{X} - \hat{\mathbf{X}}\| = \min_{\mathrm{rank}(\mathbf{Y})=r} \|\mathbf{X} - \mathbf{Y}\|. \tag{14}$$

Solution of such matrix low rank approximation problems, and indeed the total least squares problem, depends heavily on the singular value decomposition (SVD). If the Frobenius norm is chosen, then problem (14) is equivalent to problem (13).

The SVD can be described in general terms as follows. Let $\mathbf{Q} \in \mathbb{R}^{m\times n}$. Then there are orthonormal matrices $\mathbf{U} = [U_1, \ldots, U_n] \in \mathbb{R}^{m\times n}$ and $\mathbf{V} = [V_1, \ldots, V_n] \in \mathbb{R}^{n\times n}$ so that $\mathbf{Q} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{U}\,\mathrm{diag}(\sigma_1, \ldots, \sigma_n)\mathbf{V}^T$ where $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_n$. Proof of this result is included in Van Huffel and Vanderwalle (1991). If $\mathrm{rank}(\mathbf{Q}) = r$ then $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_r = \sigma_{r+1} = \cdots = \sigma_n$.

Under the assumption that $\mathrm{rank}(\mathbf{Q}) = r$, the SVD of $\mathbf{Q}$ may also be written

$$\mathbf{Q} = \sum_{i=1}^{r} \sigma_i U_i V_i^T = \sigma_1 U_1 V_1^T + \cdots + \sigma_r U_r V_r^T,$$

where $\mathrm{rank}(\sigma_i U_i V_i^T)$ for $i = 1, \ldots, r$. The Frobenius norm (as used in total least squares) of $\mathbf{Q}$ is given by

$$\|\mathbf{Q}\|_F = \sum_{i=1}^{r} \sigma_i^2.$$

The L_2 norm of $\mathbf{Q}$ is given by

$$\|\mathbf{Q}\|_2 = \sigma_1.$$

An important theorem relating the SVD to problem (14) is the so-called Eckart-Young-Mirsky matrix approximation theorem which can be formulated as follows. Assume $\mathrm{rank}(\mathbf{Q}) = r$ so $\mathbf{Q} = \sum_{i=1}^{r} \sigma_i U_i V_i^T$. Let $r' < r$ and $\mathbf{Q}' = \sum_{i=1}^{r'} \sigma_i U_i V_i^T$. Then

$$\min_{\mathrm{rank}(\mathbf{S})=\mathbf{r}'} \|\mathbf{Q} - \mathbf{S}\|_2 = \|\mathbf{Q} - \mathbf{Q}'\|_2 = \sigma_{r'+1},$$

and

$$\min_{\mathrm{rank}(\mathbf{S})=\mathbf{r}'} \|\mathbf{Q} - \mathbf{S}\|_F = \|\mathbf{Q} - \mathbf{Q}'\|_F = \sqrt{\sum_{i=r'+1}^{r} \sigma_i^2}.$$

This implies that the closest rank r approximation to Q (with respect to the L_2 norm or Frobenius norm) is given by truncating the SVD.

As given by Markovsky (2008), there are many sources of information corresponding to the total least squares topic. A detailed reference is the book by Van Huffel and Vandewalle (1991).

3.4. *Further Example*

EXAMPLE 5. Consider first the vector $(0, 1, 0, 1, 0, 1)$ mapped into the 5×2 Hankel matrix. Suppose that a rank 1 Hankel approximation to $\mathbf{X}_0$ is required. In this example, the matrix $\mathbf{X}_0$ itself and the Cadzow and HSVD approximations to $\mathbf{X}_0$ are

$$\mathbf{X}_0 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{X}_{\mathrm{Cadzow}} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{X}_{\mathrm{HSVD}} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

The HTLS method fails to give a solution; as part of the numerical calculations of the algorithm to estimate the parameters a, d, m and p of (4), $\log(0)$ appears.

Let $\epsilon = 0.01$, and instead take the matrix

$$\mathbf{X}_0 = \begin{pmatrix} \epsilon & 1 \\ 1 & \epsilon \\ \epsilon & 1 \\ 1 & \epsilon \\ \epsilon & 1 \end{pmatrix}.$$

In this case, the Cadzow, HSVD and HTLS approximations to $\mathbf{X}_0$ are

$$\mathbf{X}_{\text{Cadzow}} = \begin{pmatrix} 0.0039 & 0.0117 \\ 0.0117 & 0.0347 \\ 0.0347 & 0.1035 \\ 0.1035 & 0.3086 \\ 0.3086 & 0.9198 \end{pmatrix}, \qquad \mathbf{X}_{\text{HSVD}} = \begin{pmatrix} 0.1293 & 0.0154 \\ 0.0154 & 0.0018 \\ 0.0018 & 0.0002 \\ 0.0002 & 0 \\ 0 & 0 \end{pmatrix},$$

$$\mathbf{X}_{\text{HTLS}} = \begin{pmatrix} 0.5050 & 0.5050 \\ 0.5050 & 0.5050 \\ 0.5050 & 0.5050 \\ 0.5050 & 0.5050 \\ 0.5050 & 0.5050 \end{pmatrix}.$$

The Frobenius norms between $\mathbf{X}_0$ and these approximations are:

$$\|\mathbf{X}_0 - \mathbf{X}_{\text{Cadzow}}\|_F = 1.9357078, \qquad \|\mathbf{X}_0 - \mathbf{X}_{\text{HSVD}}\|_F = 2.225416,$$
$$\|\mathbf{X}_0 - \mathbf{X}_{\text{HTLS}}\|_F = 1.56327.$$

Now taking $\epsilon = 2$ gives the Cadzow, HSVD and HTLS approximations to $\mathbf{X}_0$ as

$$\mathbf{X}_{\text{Cadzow}} = \begin{pmatrix} 1.5629 & 1.5369 \\ 1.5369 & 1.5113 \\ 1.5113 & 1.4861 \\ 1.4861 & 1.4614 \\ 1.4614 & 1.4370 \end{pmatrix}, \qquad \mathbf{X}_{\text{HSVD}} = \begin{pmatrix} 1.5049 & 1.5031 \\ 1.5031 & 1.5012 \\ 1.5012 & 1.4994 \\ 1.4994 & 1.4975 \\ 1.4975 & 1.4957 \end{pmatrix},$$

$$\mathbf{X}_{\text{HTLS}} = \begin{pmatrix} 1.5 & 1.5 \\ 1.5 & 1.5 \\ 1.5 & 1.5 \\ 1.5 & 1.5 \\ 1.5 & 1.5 \end{pmatrix}.$$

The Frobenius norms between $\mathbf{X}_0$ and these approximations are:

$$\|\mathbf{X}_0 - \mathbf{X}_{\text{Cadzow}}\|_F = 1.577681, \qquad \|\mathbf{X}_0 - \mathbf{X}_{\text{HSVD}}\|_F = 1.580577,$$
$$\|\mathbf{X}_0 - \mathbf{X}_{\text{HTLS}}\|_F = 1.581139.$$

In this example, Cadzow algorithm finds a better approximation to the global minimum.

4. Parameterising Low Rank Matrices

For some simple examples, it is possible to visualize the space of rank deficient structured matrices by considering the determinant of the matrix to which a rank deficient approxi-

mation is required. These examples demonstrate the complexity of the space of stuctured rank deficient matrices.

EXAMPLE 6. Let $N = 3$, $L = K = 2$, $r = 1$. Define

$$H_1 = \begin{pmatrix} f_1 & f_2 \\ f_2 & f_3 \end{pmatrix} \in \mathcal{H}.$$

Then $\det(H_1) = f_1 f_3 - f_2^2$. If $\operatorname{rank}(H_1) < 2$ then it follows that $\det(H_1) = 0$. A plot of $\det(H_1) = 0$ against f_1, f_2 and f_3 is included is included in Fig. 9.

EXAMPLE 7. Let $N = 5$, $L = K = 3$, $r \leqslant 2$. Define

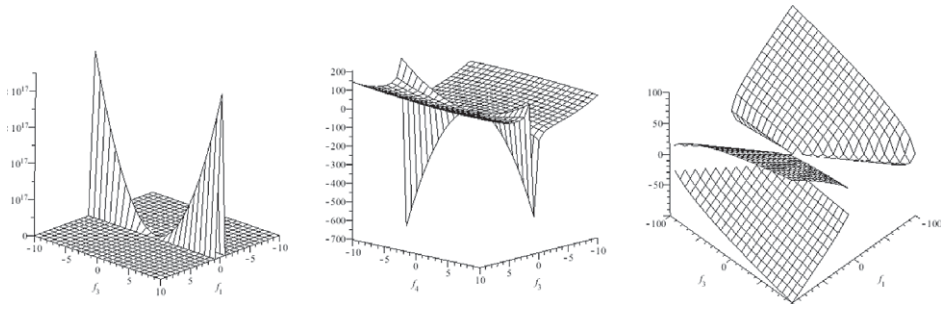
$$H_2 = \begin{pmatrix} f_1 & f_2 & f_3 \\ f_2 & f_3 & f_4 \\ f_3 & f_4 & f_5 \end{pmatrix} \in \mathcal{H}.$$

Then $\det(H_2) = f_1(f_3 f_5 - f_4^2) - f_2(f_2 f_5 - f_3 f_4) + f_3(f_2 f_4 - f_3^2)$. If $\operatorname{rank}(H_2) < 3$ then it follows that $\det(H_2) = 0$. For simplicity, setting $f_1 = f_2 = 1$ gives $\det(H_2) = f_3 f_5 - f_5 + 2 f_3 f_4 - f_4^2 - f_3^2 = 0$. A plot of $\det(H_2) = 0$ against f_3, f_4 and f_5 is included in Fig. 9.

EXAMPLE 8. This example was also considered by Chu *et al.* (2003). Let $N = 3$, $L = K = 3$, $r \leqslant 2$. Define

$$T_1 = \begin{pmatrix} f_1 & f_2 & f_3 \\ f_2 & f_1 & f_2 \\ f_3 & f_2 & f_1 \end{pmatrix} \in \mathcal{T} \subset \mathbb{R}^{3\times 3} \quad \text{(the set of Toeplitz matrices)}.$$

Then $\det(T_1) = (f_1 - f_3)(f_1^2 + f_1 f_3 - 2 f_2^2)$. If $\operatorname{rank}(T_1) < 3$ then it follows that $\det(T_1) = 0$. A plot of $\det(T_1) = 0$ against f_1, f_2 and f_3 is included in Fig. 9.



(a) Manifold of $\det(H_1) = 0$ (b) Manifold of $\det(H_2) = 0$ (c) Manifold of $\det(T_1) = 0$

Fig. 9. Determinants of rank deficient matrices.

5. Conclusion

In this paper, we have investigated the SLRA (structured low-rank approximation) problem as an optimization problem. We have shown that the optimization problem is very difficult with the objective function possessing many local minima. The objective functions has very large Lipschitz constants which increase with N, the number of observations. Additionally, the number of local minima in the neighbourhood of the global minimum increases linearly in N. Adding noise to the observed data increases the complexity of the objective function and moves the global minimizer away from the true value. We have described some algorithms that have been developed to solve the SLRA problem, using many examples to highlight the difficulty inherent in solving the problem. It is clear that the classical methods often do not even converge to a locally optimal matrix. We suggest that the difficulties described in this paper open huge possibilities for the application of the stochastic methods of g lobal optimization.

References

Abatzoglou, T., Mendel, J., Harada, G. (1991). The constrained total least squares techinique and its application to harmonic superresolution. *IEEE Transactions on Signal Processing*, 39, 1070–1087.

Broomhead, D.S., King, G.P. (1986a). Extracting qualitative dynamics from experimental data. *Physica D*, 20, 217–236.

Broomhead, D.S., King, G.P. (1986b). On the qualitative analysis of experimental dynamical systems. In: *Nonlinear Phenomena and Chaos*. Hilger, Bristol, 113–144.

Cadzow, J. (1988). Signal enhancement: a composite property mapping algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(2), 49–62.

Calvin, J.M., Žilinskas, A. (2000). On the convergence of the P-algorithm for one dimensional global optimization of smooth functions. *Journal of Optimzation Theory and Applications*, 102, 479–495.

Chu, M., Funderlic, R. and Plemmons, R. (2003). Structured low rank approximation. *Linear Algebra and Its Application*, 366, 157–172.

Clausen, J., Žilinskas, A. (2002). Subdivision, sampling and initialization strategies for simplical branch and bound in global optimization. *Computers and Mathematics with Applications*, 44(7), 943–955.

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

Elsner, J.B., Tsonis, A. A. (1996). *Singular Spectrum Analysis: A New Tool in Time Series Analysis*. Plenum, New York.

Gantmacher, F. R. (1959). *The Theory of Matrices*. Chelsea, New York.

Gillard, J. (2010). Cadzow's basic algorithm, alternating projections and singular spectrum analysis. *Statistics and Its Interface*, 3(3), 335–343.

Golyandina, N. (2010). On the choice of parameters in singular spectrum analysis and related subspace-based methods. *Statistics and Its Interface*, 3(3), 259–279.

Golyandina, N., Nektutkin, V., Zhigljavsky, A. (2001). *Analysis of Time Series Structure: SSA and related techniques*. Chapman & Hall/CRC, Boca Raton.

Hastie, T., Tibshirani, R. and Friedman, J. (1998). *The Elements of Statistical Learning*. Second edn. Springer, New York.

Lemmerling, P. and Van Huffel, S. (2001). Analysis of the structured total least squares problem for Hankel/Toeplitz matrices. *Numerical Algorithms*, 1(27), 89–114.

Markovsky, I. (2008). Structured low-rank approximation and its applications. *Automatica*, 44, 891–909.

Markovsky, I. (2010). Bibliography on total least squares and related methods. *Statistics and Its Interface*, 3(3), 329–334.

Markovsky, I., Willems, J.C., Van Huffel, S., De Moor, B. (2006). *Exact and Approximate Modeling of Linear Systems: A Behavioral Approach*. SIAM, Philadelphia.
Mockus, J. (1989). *Bayesian Approach to Global Optimization*, Mathematics and its Applications (Soviet Series). Kluwer Academic, Dordrecht.
Pollock, D.S.G. (1999). *A Handbook of Time Series Analysis, Signal Processing, and Dynamics*. Academic Press, London.
Van Huffel, S. (1993). Enhanced resolution based on minimum variance estimation and exponential data modelling. *Signal Processing*, 33(3), 333–355.
Van Huffel, S., Vandewalle, J. (1991). *The Total Least Squares Problem: Computational Aspects and Analysis*. SIAM, Philadelphia.
Zhigljavsky, A.A. (1991). *Theory of Global Random Search*. Kluwer Academic, Dordrecht.
Zhigljavsky, A., Žilinskas, A. (2008). Stochastic global optimization. *Optimization and Its Applications*, 9.

J. Gillard, PhD, is a lecturer in statistics at Cardiff University. His research interests are in time series analysis (particularly in singular spectrum analysis) and in matrix low rank optimization problems.

A. Zhigljavsky, PhD, is professor and chair of statistics at Cardiff University. He is also director of the Centre for Optimization and Its Applications. His research interests are in time series analysis, multivariate statistical analysis, statistical modelling in market research, stochastic global optimization, probabilistic methods in search and number theory and the dynamical system approach for studying convergence of search algorithms.

Struktūrinė žemo rango aproksimacija kaip optimizavimo uždavinys

Jonathan GILLARD, Anatoly ZHIGLJAVSKY

Taip vadinama struktūrinė žemo rango aproksimacija nagrinėjama kaip optimizacijos uždavinys vektorių arba matricų aibėje. Trumpai šį uždavinį galima suformuluoti šitaip. Duota pradinė specialios formos matrica (pavyzdžiui, Hankelio), ir reikia rasti ją aproksimuojančią panašiai apibrėžtą žemo rango matricą. Parodyta, kad šis uždavinys yra sunkus daugiaekstremalus uždavinys. Parodyta, kad tradiciniai metodai šiam uždaviniui netgi nekonverguoja. Nagrinėjami įvairūs metodai šiam uždaviniui spręsti.