Global optimization for structured low rank approximation

Jonathan Gillard and Anatoly Zhigljavsky

Cardiff School of Mathematics

Abstract. In this paper, we investigate the complexity of the numerical construction of the Hankel structured low-rank approximation (HSLRA) problem, and develop a family of algorithms to solve this problem. Briefly, HSLRA is the problem of finding the closest (in some pre-defined norm) rank r approximation of a given Hankel matrix, which is also of Hankel structure. Unlike many other methods described in the literature the family of algorithms we propose has the property of guaranteed convergence.

Keywords: Structured low rank approximation, Hankel matrices, global optimization
PACS: <Replace this text with PACS numbers; choose from this list: `http://www.aip..org/pacs/index.html`>

INTRODUCTION

Statement of the problem

Let L, K and r be given positive integers such that $1 \le r < L \le K$. Denote the set of all real-valued $L \times K$ matrices by $\mathbb{R}^{L\times K}$. Let $\mathscr{M}_r = \mathscr{M}_r^{L\times K} \subset \mathbb{R}^{L\times K}$ be the subset of $\mathbb{R}^{L\times K}$ containing all matrices with rank $\le r$, and $\mathscr{H} = \mathscr{H}^{L\times K} \subset \mathbb{R}^{L\times K}$ be the subset of $\mathbb{R}^{L\times K}$ containing matrices of some known structure. The set of structured $L\times K$ matrices of rank $\le r$ is $\mathscr{A} = \mathscr{M}_r \cap \mathscr{H}$. Assume we are given a matrix $\mathbf{X}_* \in \mathscr{H}$. The problem of structured low rank approximation (SLRA) is:

$$f(\mathbf{X}) \to \min_{\mathbf{X}\in\mathscr{A}} \tag{1}$$

where $f(\mathbf{X}) = \rho^2(\mathbf{X}, \mathbf{X}_*)$ is a squared distance on $\mathbb{R}^{L\times K} \times \mathbb{R}^{L\times K}$.

In this paper we only consider the case where $\mathscr{H}$ is the set of Hankel matrices and thus refer to (1) as HSLRA. Recall that a matrix $\mathbf{X} = (x_{lk}) \in \mathbb{R}^{L\times K}$ is called Hankel if $x_{lk} = \text{const}$ for all pairs (l,k) such that $l+k = \text{const}$; that is, all elements on the anti-diagonals of $\mathbf{X}$ are equal. There is a one-to-one correspondence between $L\times K$ Hankel matrices and vectors of size $N = L+K-1$. For a vector $Y = (y_1,\ldots,y_N)^T$, the matrix $\mathbf{X} = \mathbb{H}(Y) = (x_{lk}) \in \mathbb{R}^{L\times K}$ with elements $x_{lk} = y_{l+k-1}$ is Hankel and vise-versa: for any matrix $\mathbf{X} \in \mathscr{H}$, we may define $Y = \mathbb{H}^{-1}(\mathbf{X})$ so that $\mathbf{X} = \mathbb{H}(Y)$.

HSLRA is a very important problem with applications in a number of different areas. In addition to the clear connection with time series analysis and signal processing, HSLRA has been extensively used in system identification (modeling dynamical systems) [13], in speech and audio processing [11], in modal and spectral analysis [18] and image processing [15]. Some discussion on the relationship of HSLRA with some well known subspace-based methods of time series analysis and signal processing is given in [8].

There are a number of ways of parameterising the function f. One such way is via the sums of damped sinusoids:

$$f(\theta) = \sum_{n=1}^{N} (y_n - \eta(\theta,n))^2 \to \min_{\theta\in\Theta},\ \Theta \subset \mathbb{R}^n, \tag{2}$$

and the function $\eta(\theta,n)$ has the form

$$\eta(\theta,n) = \sum_{i=1}^{q} a_i \exp(d_i t)\sin(2\pi\omega_i n + \phi_i),\ n = 1,\ldots,N\ . \tag{3}$$

Here, q is a given integer, $\theta = (a,d,\omega,\phi)$ with $a = (a_1,\ldots,a_q)$, $d = (d_1,\ldots,d_q)$, $\omega = (\omega_1,\ldots,\omega_q)$ and $\phi = (\phi_1,\ldots,\phi_q)$.

We use this parameterisation (2) to make some comments about the complexity of the HSLRA problem. Objective functions are typically highly multiextremal with the objective functions possessing many local minima (see also the related discussion in [10]). Figure 1 contains some plots of (2) with (3), $q = 2$ and $N = 10$. Although the objective functions are often Lipschitz-continuous (see, e. g., [7, 10, 16, 17, 19]), it has very high Lipschitz constants which increase with N, the number of observations. Adding noise to the observed data increases the complexity of the objective function (see, e. g., [3, 20]) and moves the global minimizer away from the vector of true parameters. Thus, efficient global optimization techniques should be used to tackle the stated problem.

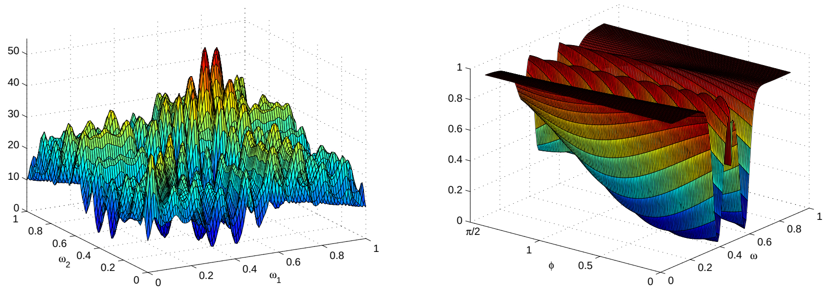

FIGURE 1. Objective functions for some small example. $f(\omega_1,\omega_2)$ (left). Cross-section (ω,ϕ) of $f(d,\omega,\phi)$ (right)

HSLRA AS AN OPTIMIZATION PROBLEM

Distances defining the objective function

There are two natural distance functions ρ which define the objective function f in (1). The most natural squared distance ρ^2 is determined by the squared Frobenius norm:

$$||\mathbf{X}||_F^2 = \sum_{l=1}^{L}\sum_{k=1}^{K} x_{lk}^2 \quad \text{for } \mathbf{X} \in \mathbb{R}^{L\times K}\,. \tag{4}$$

Every $L\times K$ Hankel matrix $\mathbf{X}\in\mathscr{H}$ is in a one-to-one correspondence with some vector $Y=(y_1,\dots,y_N)^T$, with $N=L+K-1$. Let the function $\mathbb{H}:\mathbb{R}^N\to\mathscr{H}^{L\times K}$ be defined such that $\mathbb{H}(Y)=||y_{l+k-1}||_{l,k=1}^{L,K}$ for $Y=(y_1,\dots,y_N)^T$; that is, $\mathbb{H}(Y)$ maps a vector $Y\in\mathbb{R}^N$ to an $L\times K$ Hankel matrix. Each element of the vector Y is repeated in $\mathbf{X}=\mathbb{H}(Y)$ several times. Let $\mathbf{E}=(e_{lk})\in\mathbb{R}^{L\times K}$ be the matrix consisting entirely of ones. We can compute the sum of each anti-diagonal of $\mathbf{E}$, denoted v_n, as

$$v_n = \sum_{l+k=n+1} e_{lk} = \begin{cases} n & \text{for } n=1,\dots,L-1\,, \\ L & \text{for } n=L,\dots,K-1, \\ N-n+1 & \text{for } n=K,\dots,N. \end{cases} \tag{5}$$

The value v_n is the number of times the element y_n of the vector Y is repeated in the Hankel matrix $\mathbb{H}(Y)$. Denote by $\mathbf{V}=\operatorname{diag}(v_1,\dots,v_N)$ the diagonal matrix with diagonal elements $v_1,\dots,v_N$.

If we compute the norm (4) for the Hankel matrix $\mathbf{X}=\mathbb{H}(Y)$ and express this formula in terms of the associated vector Y, then we obtain

$$||\mathbf{X}||_F^2 = \sum_{n=1}^{N} v_n y_n^2 = Y^T\mathbf{V}Y \text{ for } \mathbf{X}=\mathbb{H}(Y) \text{ with } x_{lk}=y_{l+k-1}\,. \tag{6}$$

The squared Euclidian norm of the vector Y (associated with the matrix $\mathbf{X}=\mathbb{H}(Y)$) defines another common distance ρ:

$$||\mathbf{X}||^2 = \sum_{n=1}^{N} y_n^2 = Y^TY \quad \text{for } \mathbf{X}=\mathbb{H}(Y)\,. \tag{7}$$

The general weighted squared distance is defined as

$$||\mathbf{X}||_W^2 = Y^T\mathbf{W}Y \tag{8}$$

where $\mathbf{W}$ is an arbitrary non-negative definite matrix which can sometimes be interpreted as a covariance matrix of the observations Y. For the cases $\mathbf{W}=\mathbf{V}$ and $\mathbf{W}=\mathbf{I}_N$, the squared distance (8) reduces to (6) and (7), respectively.

Projection to $\mathscr{M}_r$ for $\mathbf{W} = \mathbf{V}$ (Frobenius norm)

Let $\sigma_i = \sigma_i(\mathbf{X})$, the singular values of $\mathbf{X}$, be ordered such that $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_L$. Denote $\Sigma_0 = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_L)$ and $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r, 0, \ldots, 0)$. Then the SVD of $\mathbf{X}$ can be written as $\mathbf{X} = U\Sigma_0 V^T$ and the matrix $\pi^{(r)}(\mathbf{X}) = U\Sigma V^T$ belongs to $\mathscr{M}_r$ and minimizes the value $||\mathbf{X} - \mathbf{X}'||_F^2$ over $\mathbf{X}' \in \mathscr{M}_r$. The projection $\pi^{(r)}(\mathbf{X})$ of $\mathbf{X}$ onto $\mathscr{M}_r$ is uniquely defined if and only if $\sigma_r > \sigma_{r+1}$. The squared (Frobenius) distance between matrix $\mathbf{X}$ and $\mathscr{M}_r$ is $\rho^2(\mathbf{X}, \mathscr{M}_r) = \min\limits_{\mathbf{X}' \in \mathscr{M}_r} \rho^2(\mathbf{X}, \mathbf{X}') = \rho^2(\mathbf{X}, \pi^{(r)}(\mathbf{X})) = ||\mathbf{X} - \pi^{(r)}(\mathbf{X})||_F^2 = \sum\limits_{i=r+1}^{L} \sigma_i^2(\mathbf{X})\,.$

Projection to $\mathscr{H}$

Let $\pi_{\mathscr{H}}(\mathbf{X})$ denote the projection of a matrix $\mathbf{X} \in \mathbb{R}^{L\times K}$ onto the space $\mathscr{H}$. Then the element $\tilde{x}_{ij}$ of $\pi_{\mathscr{H}}(\mathbf{X})$ is given by $\tilde{x}_{ij} = v_{i+j}^{-1} \sum\limits_{l+k=i+j} x_{lk}$. The squared (Frobenius) distance between matrix $\mathbf{X}$ and the space $\mathscr{H}$ is $\rho^2(\mathbf{X}, \mathscr{H}) = \min\limits_{\mathbf{X}' \in \mathscr{H}} \rho^2(\mathbf{X}, \mathbf{X}') = ||\mathbf{X} - \pi_{\mathscr{H}}(\mathbf{X})||_F^2\,.$

ALGORITHMS BASED ON THE USE OF ALTERNATING PROJECTIONS

In this section we consider algorithms for solving the HSLRA problem represented as optimization problems using alternating projections between the spaces $\mathscr{H}$ and $\mathscr{M}_r$. We restrict our attention to the distance function associated with the matrix Frobenuis norm (6), that is, we take $\mathbf{W} = \mathbf{V}$ in (8).

Classical algorithms and their modifications

The algorithm (9) below is the direct implementation of the alternating projections. For brevity we will refer to this algorithm as AP.

$$\mathbf{X}_0 = \mathbf{X}_*, \quad \mathbf{X}_{n+1} = \pi_{\mathscr{H}}\left[\pi^{(r)}(\mathbf{X}_n)\right] \quad \text{for } \; n = 0, 1, \ldots \tag{9}$$

These projections have also been studied in [2] and are sometimes known as Cadzow iterations [5].

Despite AP often appearing to be myopic and too greedy by only aiming at minimizing the distance $\rho^2(\mathbf{X}, \mathscr{M}_r)$, it is very popular in practice. The popularity of AP is explained by the simplicity of the algorithm and by the fact that convergence to the space $\mathscr{A}$ is guaranteed, see [4]. AP often converges to a matrix which is far away from the set of optimal solutions $\mathfrak{X}^*$. As shown in [1, Th. 6.1], AP converges linearly; that is, there exist constants $c < 1$ and $A > 0$ such that $\rho^2(\mathbf{X}_\infty, \mathbf{X}_n) < Ac^n$, $\forall n$, where $\mathbf{X}_\infty$ is some matrix in $\mathscr{A}$. Moreover, it is easy to prove monotonicity of AP iterations. As derived by Chu et al. [4], we have $||\mathbf{X}_{n+1} - \pi^{(r)}(\mathbf{X}_{n+1})||_F^2 \leq ||\mathbf{X}_{n+1} - \pi^{(r)}(\mathbf{X}_n)||_F^2 \leq ||\mathbf{X}_n - \pi^{(r)}(\mathbf{X}_n)||_F^2\,.$

Alternating Projections with Backtracking and Randomization

In this section, we describe a family of algorithms which can be run as a random multistart-type algorithm, as a multistage algorithm and also as an evolutionary method. The main steps of this algorithm are summarized by its title ‘Alternating Projections with Backtracking and Randomization’ and we abbreviate this algorithm APBR. Here we describe two versions of this algorithm, Multistart APBR and APBR with selection. APBR with selection significantly reduces the number of computations by terminating non-prospective trajectories at early stages.

The multistart version of APBR is described as follows. Let U denote a realization of a random number with uniform distribution in $[0,1]$ and let $\tilde{\mathbf{X}}$ denote a random Hankel matrix which corresponds to a realization of a white noise Gaussian process $\tilde{Y} = (\xi_1, \ldots, \xi_N)$ with ξ_i, $i = 1, \ldots N$, independent Gaussian random variables with mean 0 and variance $s^2 \geq 0$.

In Multistart APBR, we run M independent trajectories in the space $\mathscr{H}$ starting at random Hankel matrices

$$\mathbf{X}_{0,j} = (1 - s_0)\mathbf{X}_* + s_0\tilde{\mathbf{X}}, \tag{10}$$

with some s_0 $(0 \leq s_0 \leq 1)$, and use the updating formula

$$\mathbf{X}_{n+1,j} = \left(\mathrm{tr}\mathbf{Z}_{n,j}^T\mathbf{X}_* / \mathrm{tr}\mathbf{Z}_{n,j}^T\mathbf{Z}_{n,j}\right)\mathbf{Z}_{n,j} \qquad (11)$$

where $j = 1, \ldots, M$,

$$\mathbf{Z}_{n,j} = (1-\delta_n)\,\pi_{\mathscr{H}}\left[\pi^{(r)}(\mathbf{X}_{n,j})\right] + \delta_n\mathbf{X}_* + \sigma_n\tilde{\mathbf{X}} \qquad (12)$$

and

$$\begin{cases} \delta_n = U/(n+1)^p, \quad \sigma_n = c/(n+1)^q, & \text{if } \rho^2(\mathbf{X}_{n,j}, \mathscr{M}_r) \geq \varepsilon, \\ \delta_n = 0, \; \sigma_n = 0, & \text{otherwise}. \end{cases} \qquad (13)$$

Each trajectory is either run until convergence or for a pre-specified number of iterations. U could be either random or simply set to 1, $c \in \{0,1\}$ and positive numbers p, q and ε can be chosen arbitrarily. A MATLAB implementation of this version of APBR, developed by the authors, is available at [6].

If $s_0 = \delta_n = \sigma_n = 0$ then the iterations in (11) coincide with iterations of AP (9) with some local improvement. If $s_0 > 0$ then the j-th trajectory of the algorithm starts at a random matrix in the neighbourhood of $\mathbf{X}_*$ (the width of this neighbourhood is controlled by the parameter s_0). If $\sigma_n > 0$ then there is a 'random mutation' at the n-th iteration (11). When $\delta_n > 0$, the current approximation 'backtracks' towards $\mathbf{X}_*$ conditionally that the backtracking does not worsen the distance $\rho^2(\mathbf{X}_{n,j}, \mathbf{X}_*)$. If $\rho^2(\mathbf{X}_{n,j}, \mathscr{M}_r) < \varepsilon$, we set $\delta_n = 0$ and $\sigma_n = 0$. That is, in the final stage for any trajectory of the APBR we perform AP iterations (9) to achieve faster convergence to $\mathscr{A}$.

REFERENCES

1. F. Andersson and M. Carlsson. Alternating projections on non-tangential manifolds. *arXiv:1107.4055*, 2011.
2. J. A. Cadzow. Signal enhancement: A composite property mapping algorithm. *IEEE Trans. on Acoust., Speech, Signal Processing*, 36:1070–1087, 1988.
3. J. M. Calvin and A. Žilinskas. One-dimensional global optimization for observations with noise.
4. M. T. Chu, R. E. Funderlic, and R. J. Plemmons. Structured low rank approximation. *Linear algebra and its applications*, 366:157–172, 2003.
5. J. Gillard. Cadzow's basic algorithm, alternating projections and singular spectrum analysis. *Statistics and Its Interface*, 3(3):335–343, 2010.
6. J. W. Gillard and A. A. Zhigljavsky. Software for alternating projections with backtracking and randomization, *http://www.jonathangillard.co.uk*. 2012.
7. J. W. Gillard and A. A. Zhigljavsky. Stochastic algorithms for solving structured low-rank matrix approximation problems. *Communications in Nonlinear Science and Numerical Simulation*, 21(1):70–88, 2015.
8. N. Golyandina. On the choice of parameters in Singular Spectrum Analysis and related subspace-based methods. *Statistics and Its Interface*, 3:259–279, 2010.
9. D. E. Kvasov and Y. D. Sergeyev. Lipschitz global optimization methods in control problems *Automation and Remote Control*, 74(9):1435–1448, 2013.
10. D. E. Kvasov and Y. D. Sergeyev. Deterministic approaches for solving practical black-box global optimization problems. *Advances in Engineering Software*, 80:58–66, 2015.
11. P. Lemmerling, N. Mastronardi, and S. Van Huffel. Efficient implementation of a structured total least squares based speech compression method. *Linear Algebra Appl.*, 366:295–315, 2003.
12. D. Lera and Y. D. Sergeyev. Deterministic global optimization using space-filling curves and multiple estimates of Lipschitz and Holder constants. *Communications in Nonlinear Science and Numerical Simulation.*, 23:328–342, 2015.
13. I. Markovsky, J. C. Willems, S. Van Huffel, B. De Moor, and R. Pintelon. Application of structured total least squares for system identification and model reduction. *IEEE Trans. Automat. Control*, 50(10):1490–1500, 2005.
14. R. Paulavicius, Y. D. Sergeyev, D. E. Kvasov, J. Zilinskas. Globally-biased DISIMPL algorithm for expensive global optimization. *Journal of Global Optimization.* 59(2-3):545–567, 2014.
15. A. Pruessner and D. P. O'Leary. Blind deconvolution using a regularized structured total least norm algorithm. *SIAM J. Matrix Anal. Appl.*, 24(4):1018–1037, 2003.
16. Y. D. Sergeyev and D. E. Kvasov. Lipschitz global optimization. *Wiley Encyclopedia of Operations Research and Management Science*, 2011.
17. R. G. Strongin and Y. D. Sergeyev. *Global optimization with non-convex constraints: Sequential and parallel algorithms*, volume 45. Springer Science & Business Media, 2013.
18. A. Yeredor. Multiple delays estimation for chirp signals using structured total least squares. *Linear Algebra Appl.*, 391:261–286, 2004.
19. A. Zhigljavsky and A. Žilinskas. *Stochastic Global Optimization*, volume 9. Springer Science & Business Media, 2007.
20. A. Žilinskas. On similarities between two models of global optimization: statistical models and radial basis functions. *Journal of Global Optimization*, 48(1):173–182, 2010.