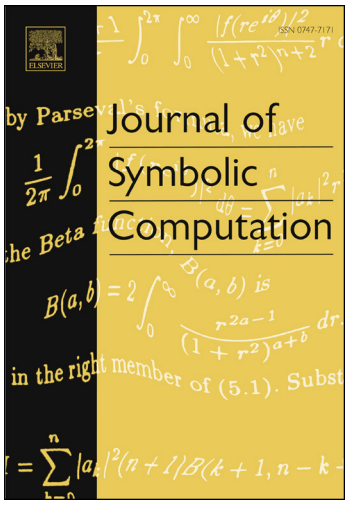
Accepted Manuscript



Computing lower rank approximations of matrix polynomials

Mark Giesbrecht, Joseph Haraldson, George Labahn

PII: S0747-7171(19)30077-X
DOI: https://doi.org/10.1016/j.jsc.2019.07.012
Reference: YJSCO 1946

To appear in: *Journal of Symbolic Computation*

Received date: 2 December 2017
Accepted date: 28 February 2019

Please cite this article in press as: Giesbrecht, M., et al. Computing lower rank approximations of matrix polynomials. *J. Symb. Comput.* (2019), https://doi.org/10.1016/j.jsc.2019.07.012

Computing Lower Rank Approximations of Matrix Polynomials[†]

Mark Giesbrecht[1], Joseph Haraldson[1], George Labahn[1]

Abstract

Given an input matrix polynomial whose coefficients are floating point numbers, we consider the problem of finding the nearest matrix polynomial which has rank at most a specified value. This generalizes the problem of finding a nearest matrix polynomial that is algebraically singular with a prescribed lower bound on the dimension given in a previous paper by the authors. In this paper we prove that such lower rank matrices at minimal distance always exist, satisfy regularity conditions, and are all isolated and surrounded by a basin of attraction of non-minimal solutions. In addition, we present an iterative algorithm which, on given input sufficiently close to a rank-at-most matrix, produces that matrix. The algorithm is efficient and is proven to converge quadratically given a sufficiently good starting point. An implementation demonstrates the effectiveness and numerical robustness of our algorithm in practice.

Keywords: Matrix polynomials; symbolic-numeric computing; low-rank approximation.

1. Introduction

Matrix polynomials appear in many areas of computational algebra, control systems theory, differential equations, and mechanics. The algebra of matrix polynomials is typically described assuming that the individual polynomial coefficients come from an exact arithmetic domain. However, in the case of applications these coefficients typically have numeric coefficients, usually real or complex numbers. As such, arithmetic can have numerical errors and algorithms are prone to numerical instability.

Numerical errors have an impact, for example, in determining the rank of a matrix polynomial with floating point coefficients. In an exact setting determining the rank or determinant of a matrix polynomial is straightforward, and efficient procedures are available, for example from Storjohann and Villard (2005). However, in a numeric environment, a matrix polynomial may appear to have full or high rank while at the same time being close to one having lower rank. Here "close" is defined naturally under the Frobenius norm on the underlying coefficient matrices of the matrix polynomial. Rather than computing the rank of the given matrix polynomial exactly, one can ask how far away it is from one that is rank-deficient, and then to find one at that distance. In the case of matrices with constant entries this is a problem solved via the Singular

[†]The authors would like the support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

[☆]Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada

Email addresses: mwg@uwaterloo.ca (Mark Giesbrecht), jharalds@uwaterloo.ca (Joseph Haraldson), glabahn@uwaterloo.ca (George Labahn)

Value Decomposition (SVD). However, in the case of matrix polynomials no equivalent rank revealing factorization has thus far been available.

In this paper we consider the problem of computing the nearest matrix polynomial to an input matrix polynomial in $\mathbb{R}[t]^{m\times n}$ having a kernel of rank at most a specified value r. More precisely, given an integer r and an $\mathcal{A} \in \mathbb{R}[t]^{m\times n}$ of full rank, we want to compute $\Delta\mathcal{A} \in \mathbb{R}[t]^{m\times n}$ with $\deg \Delta\mathcal{A}_{ij} \leq \deg \mathcal{A}_{ij}$ (or similar degree constraints to be specified later), such that $\mathcal{A} + \Delta\mathcal{A}$ has rank at most $n - r$ and where $\|\Delta\mathcal{A}\|$ is minimized. In the case where $n - r$ is one less than the row or column size then this is the problem of finding the nearest matrix polynomial which is *singular*.

A reasonable metric for measuring closeness on the space of matrix polynomials over the reals is the Frobenius norm. For a matrix polynomial $\mathcal{A} \in \mathbb{R}[t]^{m\times n}$, with (i, j) entry $A_{ij} \in \mathbb{R}[t]$, the *Frobenius* norm is given by

$$\|\mathcal{A}\|^2 = \|\mathcal{A}\|_F^2 = \sum_{1\leq i\leq m, 1\leq j\leq n} \|A_{ij}\|^2, \tag{1.1}$$

where, for a polynomial $a \in \mathbb{R}[t]$, the coefficient 2-norm is defined by

$$a = \sum_{0\leq i\leq \deg a} a_i t^i, \qquad \|a\|^2 = \|a\|_2^2 = \sum_{0\leq i\leq \deg a} a_i^2. \tag{1.2}$$

The main results in this paper center on the characterization of the geometry of minimal solutions. We show that minimal solutions exist, that is, for a given r there exists a $\Delta\mathcal{A} \in \mathbb{R}[t]^{m\times n}$ of minimal norm such that $\mathcal{A} + \Delta\mathcal{A}$ has rank at most $n - r$ and meets the required degree constraints on perturbed coefficients. In addition, we show that minimal solutions are isolated and are surrounded by a non-trivial open neighbourhood of non-minimal solutions. Also regularity and second-order sufficiency conditions are generically satisfied and a restricted version of the problem always satisfies these conditions. Finally we show that we can also generalize our results to the lower rank approximation instance of matrix polynomials generated by an affine structure[1], and so generalize to low-rank approximations of structured matrices by taking the degree to be zero.

We demonstrate efficient algorithms for computing our minimal lower rank approximants. That is, for an input matrix polynomial $\mathcal{A} \in \mathbb{R}[t]^{m\times n}$ (with prescribed affine structure) sufficiently close to a singular matrix polynomial, we give an iterative scheme which converges to a rank at most matrix polynomial at minimal distance, at a provably quadratic rate of convergence. We further generalize the iterative scheme so that it converges to a matrix polynomial with a kernel of dimension at least r, at a minimal distance and a provable quadratic rate of convergence. Finally, we also discuss a Maple implementation which demonstrates the convergence and numerical robustness of our iterative scheme.

1.1. Previous research

Much of the work in this area has often been done under the heading of *matrix pencils*. See Gohberg et al. (2009) for an excellent overview. Non-singular (full rank) square matrix polynomials are sometimes referred to as *regular matrix polynomials*.

[1] A matrix $A \in \mathbb{F}^{m\times n}$, over a ring $\mathbb{F}$, has an *affine structure* with respect to a defined set of constant matrices $\{B_0, B_1, \ldots, B_L\} \subseteq \mathbb{F}^{m\times n}$ if it can be written as $A = B_0 + \sum_{i=1}^{L} c_i B_i$ for some $c_1, \ldots, c_L \in \mathbb{F}$. If B_0 is the zero matrix, then the structure is said to be *linear*. Examples of linear structures include symmetric and hermitian matrices while matrices with an affine structure include those with entries with fixed non-zero coefficients, such as monic matrix polynomials.

In the case of finding the nearest singular matrix pencil this problem was solved by the present authors in Giesbrecht et al. (2017). Previously, this problem was posed for linear matrix pencils in Byers and Nichols (1993) and followed up in Byers et al. (1998). The nearest singular matrix polynomial relates to the stability of polynomial eigenvalue problems, linear time invariant systems and differential-algebraic equations studied subsequently in (Kressner and Voigt, 2015; Guglielmi et al., 2017). For non-linear matrix polynomials/pencils, previous works rely on embedding a non-linear (degree greater than 1) matrix polynomial into a linear matrix polynomial of much higher order. Theorem 1.1 and Section 7.2 of Gohberg et al. (2009) shows that any regular $\mathcal{A} \in \mathbb{R}[t]^{n\times n}$ of degree d can be transformed to a linear matrix polynomial $\mathcal{B} = B_0 + tB_1$, for $B_0, B_1 \in \mathbb{R}^{nd\times nd}$, which has the same non-trivial (i.e., non-one) invariant factors, and ones for the remaining invariant factor. However, this transformation is not an isomorphism (there are degree one matrix polynomials in $\mathbb{R}[t]^{nd\times nd}$ which are not the image of some degree d matrix polynomial in $\mathbb{R}[t]^{n\times n}$), nor is it distance preserving under the Frobenius norm. Hence a nearby singular matrix polynomial to $\mathcal{B} \in \mathbb{R}[t]^{nd\times nd}$ (even when constrained to a degree one perturbation) almost certainly does not correspond to a nearby singular matrix polynomial to $\mathcal{A} \in \mathbb{R}[t]^{n\times n}$. In Lawrence and Corless (2015) a more sophisticated linearization with an eye towards ameliorating this is explored.

In the context of computer algebra the notion of symbolic-numeric algorithms for polynomials has been an active area of research for a number of years, and the general framework of finding nearby instances with a desired algebraic property is being thoroughly explored. Closest to our work here is work on approximate Greatest Common Divisors (GCD) Corless et al. (1995); Beckermann and Labahn (1998b,a), multivariate polynomial factorizations Kaltofen et al. (2008), and especially the optimization-based approaches employing the Structured Total Least Norm algorithm Li et al. (2005); Kaltofen et al. (2005, 2006); Zhi (2007) and Riemannian SVD Botting et al. (2005). More recently, we have explored computing the approximate GCRD of (non-commutative) differential polynomials (Giesbrecht and Haraldson, 2014; Giesbrecht et al., 2016) and resolve similar issues.

The computer algebra community has made impressive progress on fast, exact algorithms for matrix polynomials, including nearly optimal algorithms for computing ranks, factorizations and various normal forms; see Kaltofen and Storjohann (2015) and references therein for a recent overview. Part of our goal in this current paper is establish a basis for extending the reach of these symbolic techniques to matrices of polynomials with floating point coefficients.

In a more general setting our problem can be formulated as a Structured Low Rank Approximation (SLRA) problem. A popular method to solve SLRA problems is the Structured Total Least Norm (STLN) approach (Rosen et al., 1996, 1998). These are iterative methods and in general their convergence to stationary points is linear (first order), rather than quadratic, unless additional assumptions are made. In the event STLN converges to a stationary point, there may be other stationary points arbitrarily nearby, as second order sufficient conditions may not hold. The SLRA problem is a non-linear least squares problem and accordingly other techniques such as the Restricted and Riemannian SVD (De Moor, 1993, 1994, 1995) provide general tools for solving such problems. Other heuristic tools applicable to our problem include variable projection (Golub and Pereyra, 1973, 2003) and Newton's method (Abatzoglou et al., 1991). We would expect these methods to perform very poorly in our case, as one can expect problems with large residuals to perform poorly and the rational function arising from variable projection can be too costly to deal with for modestly sized problems. The problem may also be considered as optimization on a manifold (Absil et al., 2009), however we do not explicitly consider this approach. For a detailed survey of affinely structured low-rank approximation, see (Markovsky,

2008, 2011).

Other methods for structured low-rank approximation involve the family of lift and project algorithms, with the best known being Cadzow's algorithm (Cadzow, 1988). More recently Schost and Spaenlehauer (2016) gives a sequence of alternating projections that provably converge quadratically to a fixed point. However, lift and project algorithms do not generally satisfy necessary first order (see (Bertsekas, 1999)) optimality conditions, and while they may converge (quickly) to a fixed point, there is no guarantee that the fixed point is an optimal solution, though it is usually quite good. In any case, for specific problems such as ours, understanding the geometry of the minimal solutions (and hence the well-posedness of the problem) is key to effective algorithms for their computation.

SLRA problems are in general NP-hard to solve, see for example (Poljak and Rohn, 1993; Braatz et al., 1994). They are also hard to approximate under affinely structured matrices over $\mathbb{Q}$. In general the hardness stems from determining if a bilinear system of equations admits a non-trivial solution. In the instance of classical matrix polynomials it is trivial to construct feasible points since the underlying scalar matrix problem is linearly structured.

Almost all of our contributions apply to matrix polynomials with an affine structure provided that feasible points exist, that is, singular matrix polynomials with a prescribed structure exist, which is NP-hard in general. In particular, in the degree zero case our algorithms and techniques apply to affine SLRA problems. Thus, computing the nearest (affinely structured) matrix polynomial is equivalent to SLRA problems with an affine structure.

While the contributions in this paper focus on local properties of SLRA, the local properties also imply global results. The Sum of Squares (SOS) hierarchy is a global framework for studying polynomial optimization problems subject to polynomial constraints Lasserre (2001). The SOS optimization tools have found experimental success in computing structured distances to singularity and extracting minimizers when the solutions are locally unique, see for example Henrion and Lasserre (2006). In general the SOS hierarchy converges for an infinite order of relaxations, but for several problems the relaxations converge after a finite order. The finite convergence is in polynomial time with respect to the input and the number of relaxations. In particular, this finite convergence was observed for affine SLRA problems in Henrion and Lasserre (2006) but little theory was provided to indicate the reason why. The later work of Nie (2014) shows that, under regularity and second-order sufficiency conditions, finite convergence always occurs and that it is possible to extract a minimal solution. In our contributions we prove that second-order sufficiency and regularity conditions hold generically (and if they do not, then they will hold on a restricted subset of the problem). The corollary to this is that the SOS hierarchy will have finite convergence for computing the distance of the nearest rank-deficient matrix polynomial, and if the embedding is minimal then a minimizer may be extracted as well. Another useful feature of the SOS hierarchy is even if convergence cannot be certified, a structured lower-bound is obtained.

1.2. Outline

In Sections 2 and 3 we describe tools needed for our constructions and then explore the geometry of our problem. We show that the problem is locally well-posed. One cannot expect the nearest rank at most matrix polynomial to be unique. However under weak normalization assumptions, we show that solutions are locally unique in a closed-ball around them. To complement the separation of solutions, we also show that for an equivalent problem, solutions corresponding to a different closed ball are separated by at least a constant amount independent of the dimension of the space.

In Section 4 we give an equality constrained variant of Newtons' method for computing via post-refinement the nearest rank at most matrix polynomial. The main idea is to compute an initial guess with a suitable first order or lift-and project method. We are able to prove that, with a suitable initial guess and regularity assumptions, our algorithm generally has local quadratic convergence except for degenerate cases. This is done by deriving closed-form expressions for the Jacobian of the constraints and the Hessian of the Lagrangian. When we refer to the speed of convergence, we refer to quotient rates as is typical in the nomenclature.

In Section 5 we describe our prototype implementation, including heuristics for starting points and other improvements. We discuss the numerical performance of the algorithm and give examples demonstrating convergence. results for a low-rank approximation of matrix polynomials. The paper ends with a conclusion and topics for future research.

2. Preliminaries and Geometry

In this section we will introduce some basic definitions and explore the numerical geometry of our lower rank problem. Canonically we will let

$$\mathcal{A} = \sum_{j=0}^{d} A_j t^j \in \mathbb{R}[t]^{n\times n}$$

be a matrix polynomial, with coefficients $A_0, \ldots, A_d \in \mathbb{R}^{n\times n}$. In the case of rectangular matrix polynomials we can pad the matrix with zeros, thus embedding the problem into one with square matrix polynomials. Since we are finding matrices of a prescribed reduced rank (which is presumably less than the row and column order), this does not affect the low rank approximation (in particular, there is negative benefit to introducing non-zeros into the newly introduced zero rows or columns). The *degree* $\deg \mathcal{A}$ of $\mathcal{A}$ is defined as d, assuming that $\mathcal{A}_d \neq 0$.

We say that $\mathcal{A}$ is singular if $\det(\mathcal{A})$ is the zero polynomial in $\mathbb{R}[t]$, or equivalently, that there is a $\mathcal{b} \in \mathbb{R}[t]^{n\times 1}$ such that $\mathcal{A}\mathcal{b} \equiv 0$. The kernel of $\mathcal{A}$ is $\ker \mathcal{A} = \{\mathcal{b} \in \mathbb{R}[t]^{n\times 1} : \mathcal{A}\mathcal{b} = 0\}$ and the rank of $\mathcal{A}$ is $n - \dim \ker \mathcal{A}$ (as a vector space over $\mathbb{R}(t)$). Then $\mathcal{A}$ has rank at most $n - r$ if there exists at least r linear independent vectors $\{\mathcal{b}_i\}_{i=1,\ldots,r}$ satisfying $\mathcal{A}\mathcal{b}_i = 0$.

For $a \in \mathbb{R}[t]$, define

$$\phi(a) = \phi^{(n,d)}(a) = \begin{pmatrix} a_0 & & & \\ a_1 & a_0 & & \\ \vdots & & \ddots & \\ a_d & & & a_0 \\ & a_d & & a_1 \\ & & \ddots & \vdots \\ & & & a_d \end{pmatrix} \in \mathbb{R}^{(\mu+d)\times\mu}, \tag{2.1}$$

where $\mu = nd + 1$. $\phi(a)$ is a Toeplitz matrix. Such matrices are conveniently used to describe polynomial multiplication in the sense that if $c = a \cdot b$ with a of degree d and $c \in \mathbb{R}[t]$ of degree at most $\mu - 1$, then $\mathrm{vec}(c) = \phi(a) \cdot \mathrm{vec}(b)$ where $\mathrm{vec}(p)$ is the vector of coefficients of a polynomial.

Definition 2.1. *The* $\mathbb{R}$-embedding *of* $\mathcal{A} \in \mathbb{R}[t]^{n\times n}$ *is*

$$\widehat{\mathcal{A}} = \begin{pmatrix} \phi(A_{1,1}) & \cdots & \phi(A_{1,n}) \\ \vdots & & \vdots \\ \phi(A_{n,1}) & \cdots & \phi(A_{n,n}) \end{pmatrix} \in \mathbb{R}^{n(\mu+d)\times n\mu}.$$

For $\mathcal{b} \in \mathbb{R}[t]^{n\times 1}$ *of degree* $\mu - 1$ *the* $\mathbb{R}$*-embedding of* $\mathcal{b}$ *is*

$$\widehat{\mathcal{b}} = (b_{1,0}, b_{1,1}, \ldots, b_{1,\mu-1}, \ldots, b_{n,0}, \ldots, b_{n,\mu-1})^T \in \mathbb{R}^{n\mu\times 1}.$$

Note that $\mathcal{A} \cdot \mathcal{b} = 0$, for $\mathcal{b} \in \mathbb{R}[t]$ of degree at most $\mu - 1$ if and only if $\widehat{\mathcal{A}} \cdot \widehat{\mathcal{b}} = 0 \in \mathbb{R}^{n\mu\times 1}$. This property is central to our work in the coming sections.

For ease of notation we will take

$$N = n(\mu + d) = n^2 d + n(d + 1), \quad M = n\mu = n^2 d + n \text{ and } R \geq 1$$

when dealing with $\mathbb{R}$-embeddings in subsequent sections. We note that $\widehat{\mathcal{A}}$ is a block-Toeplitz matrix, and as such one method of understanding the lower rank problem is to find close by structured rank deficient block-Toeplitz matrices, a typical structured low rank approximation problem. Some authors refer to such embeddings as a (permuted) Sylvester matrix associated with $\mathcal{A}$. We avoid this terminology as it is ambiguous when considering Sylvester matrices occurring in (approximate) GCD computations.

Unlike the standard linearizations in (Gohberg et al., 2009, Section 7.2) used to turn arbitrary degree matrix pencils into linear pencils, this $\mathbb{R}$-embedding is kernel preserving for matrix polynomials of arbitrary degree. In particular, $\mathcal{b} \in \ker \mathcal{A}$ with $\deg \mathcal{b} \leq \mu$ implies $\widehat{\mathcal{b}} \in \ker \widehat{\mathcal{A}}$. The $\mathbb{R}$-embedding is also quasi-distance preserving, since $\|\mathcal{A}\|_F^2 = \frac{\|\widehat{\mathcal{A}}\|_F^2}{\mu}$.

Problem 2.2. ***Main Problem:***
Given $\mathcal{A} \in \mathbb{R}[t]^{n\times n}$ *non-singular of degree* d *and an integer* $r \leq n - 1$*, determine* $\Delta\mathcal{A} \in \mathbb{R}[t]^{n\times n}$*, with* $\deg \Delta\mathcal{A}_{ij} \leq \deg \mathcal{A}_{ij}$ *for all* $1 \leq i, j \leq n$*, and* r *linearly independent vectors* $\mathcal{b}_k \in \mathbb{R}[t]^{n\times 1}$*, such that* $\|\Delta\mathcal{A}\|$ *is (locally) minimized, subject to the constraint that* $(\mathcal{A} + \Delta\mathcal{A})\mathcal{b}_k = 0$ *and* $\|\mathcal{b}_k\| = 1$.

Note that this is minimizing a convex objective function subject to non-convex constraints. However, the equality constraints are linear in each argument. It is still not clear that Problem 2.2 is well-posed in the current form. We will prove that solutions exist, that is, there is an attainable global minimum value and not an infimum.

Lemma 2.3. $\mathcal{A} \in \mathbb{R}[t]^{n\times n}$ *is singular if and only if there exists a* $\mathcal{b} \in \mathbb{R}[t]^{n\times 1}$ *with* $\deg \mathcal{b} \leq nd = \mu - 1$ *such that* $\mathcal{A}\mathcal{b} = 0$.

Proof. Suppose that $\mathcal{A}$ has rank $s < n$. By permuting rows and columns we may assume without loss of generality that the leading $s \times s$ submatrix of $\mathcal{A}$ is non-singular. There is a unique vector of the form

$$c = (b_1/\gamma, \ldots, b_s/\gamma, -1, 0, \ldots, 0)$$

from Cramer's rule such that $\mathcal{A}c = 0$, where $\gamma \in \mathbb{R}[t]$ is the determinant of the leading $s \times s$ minor of $\mathcal{A}$, and all of $b_1, \ldots, b_s, \gamma \in \mathbb{R}[t]$ have degree at most $sd \leq nd$. Multiplying through by γ, we find that $\mathcal{b} = \gamma c$ satisfies the requirements of the lemma. □

See (Beckermann et al., 2006, Corollary 5.5) for an alternative proof.

Lemma 2.4. *$\mathcal{A}$ is singular if and only if $\widehat{\mathcal{A}}$ does not have full column rank.*

Proof. If $\mathcal{A}$ is rank deficient then there exists $\mathcal{b} \in \mathbb{R}[t]^{n\times 1}$ with $\deg \mathcal{b} \leq \mu - 1$ such that $\mathcal{A}\mathcal{b} = 0$. $\widehat{\mathcal{A}}$ has a non-trivial kernel and, $\widehat{\mathcal{b}} \in \ker \widehat{\mathcal{A}}$ by construction. Conversely, suppose that $\mathcal{A}$ has full rank. Then for all $\mathcal{b} \in \mathbb{R}[t]^{n\times 1}$ we have $\mathcal{A}\mathcal{b} \neq 0$ which implies that $\widehat{\mathcal{A}}\widehat{\mathcal{b}} \neq 0$ or $\ker \widehat{\mathcal{A}}$ is trivial. □

We recall the Singular Value Decomposition as the primary tool for finding the distance to the nearest *unstructured* rank deficient matrix over $\mathbb{R}$ or $\mathbb{C}$.

Definition 2.5. *A Singular Value Decomposition (SVD) of $C \in \mathbb{R}^{N\times M}$ is given by $C = Q \cdot \Sigma \cdot P^T$, where $Q \in \mathbb{R}^{M\times M}, P^T \in \mathbb{R}^{N\times N}$ are orthogonal matrices and $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_M) \in \mathbb{R}^{M\times N}$ is a diagonal matrix consisting of the singular values of C in descending order of magnitude. See (Golub and Van Loan, 2012).*

The following fact is a standard motivation for the SVD.

Fact 2.6 (Eckart and Young (1936))**.** *Suppose $C = Q\Sigma P^T \in \mathbb{R}^{N\times M}$ as above has full column rank, with $N \geq M$. Then $\Delta C = Q \operatorname{diag}(0, \ldots, 0, -\sigma_M)P^T$ is such that $C + \Delta C$ has column rank at most $M - 1$, $\|\Delta C\|_F = \sigma_M$, and ΔC is a perturbation of minimal Frobenius norm which reduces the column rank of C.*

Lemma 2.7. *Given a non-singular $\mathcal{A} \in \mathbb{R}[t]^{n\times n}$, and $\Delta\mathcal{A} \in \mathbb{R}[t]^{n\times n}$ such that $\mathcal{B} = \mathcal{A} + \Delta\mathcal{A}$ is singular, it is the case that $\|\widehat{\Delta\mathcal{A}}\| \geq \sigma_{n\mu}(\widehat{\mathcal{A}})$.*

Proof. By Lemma 2.4 above, $\widehat{\mathcal{B}}$ is not of full column rank. Thus, by Fact 2.6 $\|\widehat{\Delta\mathcal{A}}\|_F \geq \sigma_{n\mu}(\mathcal{A})$. □

It follows immediately that the set of all matrices of rank at-most $n - r$ over $\mathbb{R}[t]^{n\times n}$ of degree at most d is closed.

Theorem 2.8 (Existence of Solutions)**.** *The minimization posed in Problem 2.2 has an attainable global minimum if $\deg \Delta\mathcal{A}_{i,j} \leq \deg \mathcal{A}_{i,j}$ for all $1 \leq i, j \leq n$.*

Proof. Let

$$S = \{\mathcal{C} \in \mathbb{R}[t]^{n\times n} \mid \operatorname{rank} \mathcal{C} \leq n - r \wedge \deg \mathcal{C} \leq d\}$$
$$\cap \left\{\mathcal{C} \in \mathbb{R}[t]^{n\times n} \| \|\mathcal{C}\|_F^2 \leq \|\mathcal{A}\|_F^2\right\}.$$

S is the intersection of a closed and bounded set and a closed set, hence S is closed and bounded. S is isomorphic to some closed and bounded subset of Euclidean space, hence by the Heine-Borel theorem, S is compact. To show the set is non-empty, we note that, by the degree assumption on $\Delta\mathcal{A}$, $\Delta\mathcal{A} = -\mathcal{A}$ is a feasible point independent of rank.

Let $\mathcal{C} \in S$ then $\|\mathcal{A} - \mathcal{C}\|_F^2 = \|\Delta\mathcal{A}\|_F^2$ is a continuous function over a compact set. By Weierstrass' theorem it has an attainable global minimum. □

It is important not to over-constrain the problem with a choice of $\Delta\mathcal{A}$, since otherwise the feasible set might be empty. Another reasonable choice of $\Delta\mathcal{A}$ which we can handle, is that the perturbation has the same coefficient structure/support as $\mathcal{A}$, that is, zero terms in polynomial entries are preserved.

We note that this result says nothing about uniqueness or separation of solutions or any local properties. All that has been shown is that if the perturbations are in the same space as the input, and one seeks a rank at-most approximation, then there is an attainable global minimum value, i.e. not an infimum. If one wants a minimal solution with the rank being exactly $n - r$, then there is no guarantee that there is an attainable global minimum to Problem 2.2.

3. Rank Factorization

A natural formulation of the problem that encompasses the rank implicitly is to perform a rank factorization and write $\mathcal{A} + \Delta\mathcal{A} = \mathcal{U}\mathcal{V}$ for $\mathcal{U} \in \mathbb{R}[t]^{n\times(n-r)}$ and $\mathcal{V} \in \mathbb{R}(t)^{(n-r)\times n}$. Here $\mathcal{U}\mathcal{V}$ is subject to some constraints that preserve the structure of $\Delta\mathcal{A}$ (i.e., that we do not perturb any coefficients we are not allowed to, typically that $\deg \Delta\mathcal{A}_{ij} \leq \deg \mathcal{A}_{ij}$, but possibly also preserving the zero coefficients and not introducing a larger support). This is a non-linear least squares problem. However solutions are not unique. Indeed, if $\mathcal{Z} \in \mathbb{R}[t]^{(n-r)\times(n-r)}$ is unimodular (i.e., $\det(\mathcal{Z}) \in \mathbb{R}^*$), then $\mathcal{U}\mathcal{Z}$, $\mathcal{Z}^{-1}\mathcal{V}$ is another rank $n - r$ factorization, and we obtain an infinite family. While normalizing over matrix polynomial rank-factorizations is difficult, it is much easier to exploit the quasi-distance preserving property of $\|\cdot\|_F$ and look at rank-factorizations of $\widehat{\mathcal{A}}$, that do not necessarily correspond to $\mathcal{U}$ and $\mathcal{V}$.

3.1. Embedded Rank Factorization

Definition 3.1 (Rank Factorization)**.** *Let $N = (\mu + d)n$, $M = n\mu$ and $R > 0$. A rank factorization of $\widehat{\mathcal{A}} + \widehat{\Delta\mathcal{A}}$ is given by writing $\widehat{\mathcal{A}} + \widehat{\Delta\mathcal{A}} = UV$ where $U \in \mathbb{R}^{N\times R}$ and $V \in \mathbb{R}^{R\times M}$ are arbitrary (unstructured) matrices over $\mathbb{R}$.*

Our goal is to find U, V with appropriate dimensions which minimize

$$\|\Delta\widehat{\mathcal{A}}\| = \|\widehat{\mathcal{A}} - UV\|$$

and such that $\Delta\widehat{\mathcal{A}}$ has the correct Toeplitz-block structure (i.e., it is an $\mathbb{R}$-embedding of a matrix polynomial). This is a problem with a non-convex objective function (that is convex in each argument) and non-convex constraints. We note that U, V have no direct connection with $\mathcal{U}$ and $\mathcal{V}$ discussed earlier.

One may always write $\widehat{\mathcal{A}} + \widehat{\Delta\mathcal{A}}$ this way via the SVD for fixed $\widehat{\mathcal{A}}$ and $\widehat{\Delta\mathcal{A}}$, so in particular the optimal solution can be written as a rank factorization. The problem $\min \|\widehat{\mathcal{A}} - UV\|^2$ such that UV has the same structure as $\widehat{\Delta\mathcal{A}}$ is generally ill-posed and needs to be constrained to do any meaningful analysis, as there are numerous degrees of freedom. At first glance, optimizing over rank factorizations appears to be a harder problem than the original. However it is helpful to perform analysis on this formulation. In particular, we are able to prove that optimal values of $\widehat{\Delta\mathcal{A}}$ that satisfy first order conditions (which contains all useful perturbations) are separated by a constant amount, and that equivalence classes of solutions are isolated. Additionally, this formulation of the problem is convex in each argument (but not jointly convex) and is amenable to block coordinate descent methods.

We next need to demonstrate that the condition that the matrix $\Delta\widehat{\mathcal{A}} = \widehat{\mathcal{A}} - UV$ is the $\mathbb{R}$-embedding of some matrix polynomial $\Delta\mathcal{A} \in \mathbb{R}[t]^{n\times n}$ can be phrased as a single polynomial being zero. $\widehat{\mathcal{A}}$ is generated by a linear structure $\sum_{i=1}^{L} c_i\widehat{\mathcal{A}}^{(i)}$ where $c_i \in \mathbb{R}$ and $\{\widehat{\mathcal{A}}^{(1)}, \ldots, \widehat{\mathcal{A}}^{(L)}\} \subseteq \mathbb{R}^{N\times M}$. Define the structural enforcement function

$$\Gamma : \mathbb{R}^{N\times R} \times \mathbb{R}^{R\times M} \to \mathbb{R} \text{ as } \Gamma(U,V) = \left\| \sum_{i=1}^{L} c_i\widehat{\mathcal{A}}^{(i)} - \Delta\widehat{\mathcal{A}} \right\|_F^2.$$

We note that there exist c_i such that $\Gamma(\Delta\widehat{\mathcal{A}}) = 0$ if and only if $\Delta\widehat{\mathcal{A}}$ is an $\mathbb{R}$-embedding of a matrix polynomial.

Problem 3.2. *With $\widehat{\mathcal{A}}, U, V$ as above, the constrained $\mathbb{R}$-embedded rank factorization problem consists of computing* $\min \|\widehat{\mathcal{A}} - UV\|_F^2$ *subject to the constraints that* $U^TU - I = 0$ *and* $\Gamma(U,V) = 0$*. If $R = M - 1$, then this encodes all rank deficient matrix polynomials.*

It is still not clear that Problem 3.2 is well-posed, as there are many degrees of freedom in V, and this matrix can have arbitrary rank. The enforcement of U as an orthogonal matrix ($U^TU - I = 0$) is allowed for without loss of generality. Informally then we are looking at all rank factorizations where U is orthogonal and $\Gamma(U,V) = 0$, that is, the product satisfies the block-Toeplitz structure on $\widehat{\Delta\mathcal{A}}$.

We employ the machinery of non-linear optimization to describe the geometry of the minimal solutions, and hence the nearest appropriately structured matrices. See (Bertsekas, 1999) for an excellent overview.

Fact 3.3 (Bertsekas (1999, Section 3.1.1))**.** *For a sufficiently large $\rho > 0$, one has that*[2] *Problem 3.2 is equivalent to computing a solution to the unconstrained optimization problem*

$$\Phi(U,V) = \min_{U,V} \|\widehat{\mathcal{A}} - UV\|_F^2 + \rho\|\Gamma(U,V)\|_F^2 + \rho\|U^TU - I\|_F^2.$$

All the interesting solutions to the minimization of $\Phi(U,V)$ occur at stationary points. The first-order necessary condition (on V) of gradients vanishing gives us (slightly abusing notation)

$$\begin{aligned} &\nabla_V \left(\|\widehat{\mathcal{A}} - UV\|_F^2 + \rho\|\Gamma(U,V)\|_F^2 + \rho\|U^TU - I\|_F^2 \right) = 0 \\ \iff\ & U^T(\widehat{\mathcal{A}} - UV) + \left(\frac{\partial}{\partial V}\Gamma(U,V)^T \right) \rho\Gamma(U,V) = 0. \end{aligned}$$

If we assume that the constraints are active, that is U is orthogonal and that $\Gamma(U,V) = 0$, then we have $U^T\widehat{\mathcal{A}} - V = 0$. Of course, there is the other first order necessary condition requiring that

$$\nabla_U \left(\|\widehat{\mathcal{A}} - UV\|^2 + \rho\|\Gamma(U,V)\|^2 + \rho\|U^TU - I\|^2 \right) = 0.$$

However, we do not need to employ this explicitly in the following.

Theorem 3.4 (Strong Separation of Objective)**.** *Suppose $\widehat{\Delta\mathcal{A}}$ and $\widehat{\Delta\mathcal{A}}^\star$ are distinct (local) optimal solutions to Problem 3.2 that satisfy first order necessary conditions. Then $\|\widehat{\Delta\mathcal{A}} - \widehat{\Delta\mathcal{A}}^\star\|_2 \geq \sigma_{\min}(\widehat{\mathcal{A}})$, where $\sigma_{\min}(\cdot)$ is the smallest non-trivial singular value.*

[2] ρ is sometimes known as a penalty term.

Proof. From the previously discussed necessary first order condition we have that there exists $U \in \mathbb{R}^{N\times R}$, $V \in \mathbb{R}^{R\times M}$ and $U^\star \in \mathbb{R}^{N\times R^\star}$ and $V^\star \in \mathbb{R}^{R^\star\times M}$ such that

$$\|\widehat{\Delta\mathcal{A}} - \widehat{\Delta\mathcal{A}}^\star\|_2 = \|UV - U^\star V^\star\|_2 = \|UU^T\widehat{\mathcal{A}} - U^\star U^{\star T}\widehat{\mathcal{A}}\|_2.$$

Note that R and $R^\star$ need not be the same. From this we obtain the sequence of lower bounds

$$\begin{aligned}\|UU^T\widehat{\mathcal{A}} - U^\star U^{\star T}\widehat{\mathcal{A}}\|_2 &\geq \|UU^T - U^\star U^{\star T}\|_2\sigma_{\min}(\widehat{\mathcal{A}}) \\ &= \|I - U^T U^\star U^{\star T}U\|_2\sigma_{\min}(\widehat{\mathcal{A}}) \\ &\geq \sigma_{\min}(\widehat{\mathcal{A}}).\end{aligned}$$

The symmetric matrix $W = U^T U^\star U^{\star T}U$ is a product of matrices whose non-zero eigenvalues have magnitude 1. Symmetric matrices have real eigenvalues, and the non-zero eigenvalues of W will be ± 1, since U and $U^\star$ are orthogonal. Thus $\|W\|_2 = 1$.

W must have at least one negative eigenvalue or zero eigenvalue by the orthogonality assumption, since $W \neq I$. Since W is symmetric, we can diagonalize W as a matrix with ± 1 and 0 entries on the diagonal. It follows that $\|I - W\|_2 \geq 1$ and the theorem follows. □

While the separation bound exploited properties of the rank factorization, these bounds hold for all formulations of the problem.

Corollary 3.5. *All locally optimal solutions satisfying first order necessary conditions are isolated modulo equivalence classes.*

Proof. Suppose the contrary, that is that (U, V) is a solution corresponding to $\widehat{\Delta\mathcal{A}}$ and $(U^\star, V^\star)$ is a solution corresponding to $\widehat{\Delta\mathcal{A}}^\star$. The objective function and constraints are locally Lipschitz continuous, so let $s > 0$ be a Lipschitz constant with respect to $\|\cdot\|_F$ in some open neighborhood.

If we take $0 < \varepsilon < \dfrac{\sigma_{\min}(\widehat{\mathcal{A}})}{s}$ such that $\left\|\begin{pmatrix}U\\V\end{pmatrix} - \begin{pmatrix}U^\star\\V^\star\end{pmatrix}\right\|_F < \varepsilon$ then we have that

$$\begin{aligned}\sigma_{\min}(\widehat{\mathcal{A}}) &\leq \|\Delta\widehat{\mathcal{A}} - \Delta\widehat{\mathcal{A}}^\star\|_2 \\ &\leq s\left\|\begin{pmatrix}U\\V\end{pmatrix} - \begin{pmatrix}U^\star\\V^\star\end{pmatrix}\right\|_F \\ &\leq s\varepsilon \\ &< \sigma_{\min}(\widehat{\mathcal{A}}),\end{aligned}$$

which is a contradiction to Theorem 3.4. □

Implicitly the matrix V parametrizes the kernel of $\widehat{\mathcal{A}}$. If we normalize the kernel of $\widehat{\mathcal{A}}$ to contain $\mathbb{R}$-embeddings of primitive kernel vectors then the matrix V can be made locally unique, although we do not employ this in the rank-factorization formulation directly.

Corollary 3.6. *Under a suitable choice of* $\|\cdot\|$ *we have that minimal solutions are separated. In particular, separation holds for* $\|\cdot\|_1$.

The proof follows immediately from equivalence of matrix norms, as norms are equivalent in a finite dimensional space.

While there are too many degrees of freedom to easily obtain a (locally) quadratically convergent minimization over the rank factorization, the rank factorization does yield non-trivial insights into the geometry of the solution space. In particular, the isolation of solutions indicates first order (gradient) methods will perform well on the problem. In the next section we will introduce a locally quadratically convergent algorithm for an equivalent form of Problem 2.2 that reduces each equivalence class of solutions to a single solution.

4. An Iterative Algorithm for Lower Rank Approximation

In this section we propose an iterative algorithm to solve Problem 2.2 based on Newton's method for constrained optimization. Sufficient conditions for quadratic convergence are that the second-order sufficiency holds (Wright, 2005) and local Lipschitz continuity of the objective and constraints. We ensure these conditions hold for non-degenerate problems by working on a restricted space of minimal $\mathbb{R}$-embeddings that remove degrees of freedom.

4.1. Minimal System of Equations

In order to compute a nearby rank $n - r$ approximation, we want to solve the non-convex optimization problem

$$\min \|\Delta\mathcal{A}\|_F^2 \text{ subject to } \begin{cases} (\mathcal{A} + \Delta\mathcal{A})\mathcal{B} = 0, \\ \operatorname{rank}(\mathcal{B}) = r. \end{cases} \tag{4.1}$$

In the instance of (structured) scalar matrices the rank constraint can be enforced by ensuring that $\mathcal{B}$ has orthogonal columns[3] or is in a column reduced echelon form. In the instance of matrix polynomials this is not sufficient, since polynomial multiples of the same vector will have linearly independent combined coefficient vectors. In order to apply these normalizations on the coefficient vectors of $\mathcal{B}$ we require that the columns be represented with a minimal number of equations with respect to $\mathcal{B}$.

Definition 4.1 (Minimal $\mathbb{R}$-Embedding)**.** *Suppose $\mathcal{A} \in \mathbb{R}[t]^{n\times n}$ with $\mathbb{R}$-embedding $\widehat{\mathcal{A}}$. The vector $\mathfrak{b} \in \mathbb{R}[t]^{n\times 1}$, with $\mathbb{R}$-embedding $\widehat{\mathfrak{b}}$, is said to be* minimally $\mathbb{R}$-embedded *in $\widehat{\mathcal{A}}$ if* $\ker \widehat{\mathcal{A}} = \langle\widehat{\mathfrak{b}}\rangle$ *(i.e., a dimension 1 subspace). We say that $\widehat{\mathfrak{b}}$ is* minimally degree $\mathbb{R}$-embedded *in $\widehat{\mathcal{A}}$ if (1) $\widehat{\mathfrak{b}}$ is minimally $\mathbb{R}$-embedded in $\widehat{\mathcal{A}}$ and (2) $\widehat{\mathfrak{b}}$ corresponds to a primitive kernel vector $\mathfrak{b}$, that is* $\gcd(\mathfrak{b}_1, \ldots, \mathfrak{b}_n) = 1$.

We note that this definition ensures minimally $\mathbb{R}$-embedded vectors are unique (up to scaling a factor), or that $(\widehat{\mathcal{A}}_j + \widehat{\Delta\mathcal{A}}_j)\widehat{\mathcal{B}}[*, j] = 0$ has a (locally) unique solution for fixed $\widehat{\Delta\mathcal{A}}$ (where $\widehat{\mathcal{B}}[*, j]$ is the jth column of $\widehat{\mathcal{B}}$). In the minimal embedding, we will assume, without loss of generality, that redundant or equations known in advance, such as $0 = 0, \Delta\widehat{\mathcal{A}}_{ij} = 0$ or $\widehat{\mathcal{B}}_{ij} = 0$ corresponding to known entries are removed for some indices of i and j. If we assume that $\mathcal{B}$ is primitive and $\widehat{\mathcal{B}}$ is in Column Reduced Echelon Form (CREF), then this will satisfy the minimal

[3] This normalization alone is not sufficient for rapid convergence.

embedding requirements. Some of these trivial equations occur because of the CREF (or other) assumption, while others occur from over-estimating degrees of entries.

This allows us to reformulate $(\mathcal{A} + \Delta\mathcal{A})\mathcal{B} = 0$ as a (bi-linear) system of equations

$$\{(\widehat{\mathcal{A}}_j + \Delta\widehat{\mathcal{A}}_j)\widehat{\mathcal{B}}[*, j] = 0\}_{j=1}^{r} \tag{4.2}$$

where the j^{th} column of $\mathcal{B}$ is minimally degree embedded in the system $(\widehat{\mathcal{A}}_j + \Delta\widehat{\mathcal{A}}_j)$. We also note that assuming $\mathcal{B}$ is in a column-reduced echelon form essentially requires us to guess the pivots in advance of the optimal solution, which is only possible with a good initial guess. The benefit of this approach is that if the pivots are not guessed correctly, we are still able to compute a $n - r$ approximation of $\mathcal{A}$.

In order to exclude trivial solutions, we can assume that the pivot elements of $\mathcal{B}$ have a norm bounded away from zero. Let $n(\widehat{b}_i)$ be a normalization vector such that $n(\widehat{b}_i)^T\widehat{b}_i = 1$ which implies that the CREF pivots are bounded away from zero. For example, take the pivot to have unit norm. Note that other normalization vectors are possible, such as $n(\widehat{b}_i) = \widehat{b}_i$ (which corresponds to each column having a unit norm) if the initial guess is adequately close, or we could take the pivot element to be a monic polynomial. Of course there are several other permissible normalizations.

Define the matrix $\widehat{\mathcal{A}}_i$ to have the column $\widehat{b}_i = \widehat{\mathcal{B}}[1..n, i]$ minimally degree embedded. We can express (4.2) in a vector-matrix form as follows.

$$\left(\begin{array}{cccc} \widehat{\mathcal{A}}_1 + \widehat{\Delta\mathcal{A}}_1 & & & \\ & \widehat{\mathcal{A}}_2 + \widehat{\Delta\mathcal{A}}_2 & & \\ & & \ddots & \\ & & & \widehat{\mathcal{A}}_r + \widehat{\Delta\mathcal{A}}_r \\ \hline n(\widehat{b}_1)^T & & & \\ & n(\widehat{b}_2)^T & & \\ & & \ddots & \\ & & & n(\widehat{b}_r)^T \end{array}\right) \left(\begin{array}{c} \widehat{b}_1 \\ \widehat{b}_2 \\ \vdots \\ \widehat{b}_r \end{array}\right) = \left(\begin{array}{c} 0 \\ 0 \\ \vdots \\ 0 \\ \hline 1 \\ 1 \\ \vdots \\ 1 \end{array}\right) \tag{4.3}$$

has a (locally) unique solution for fixed $\Delta\mathcal{A}$.

4.2. Lagrange Multipliers and Optimality Conditions

In order to solve (4.1) we will use the method of Lagrange multipliers (Bertsekas, 1999).

Let $M(\Delta\mathcal{A}, \mathcal{B})$ be the vector of residuals corresponding to (4.3), then the Lagrangian is defined as

$$L = \|\Delta\mathcal{A}\|_F^2 + \lambda^T M(\Delta\mathcal{A}, \mathcal{B}), \tag{4.4}$$

where $\lambda = (\lambda_1, \ldots, \lambda_{\#\text{ residuals}})^T$ is a vector of Lagrange multipliers.

Definition 4.2. *The vectorization of $\mathcal{A} \in \mathbb{R}[t]^{n\times n}$ of degree at most d is defined as*

$$\text{vec}(\mathcal{A}) = (\mathcal{A}_{1,1,0}, \ldots, \mathcal{A}_{1,1,d}, \mathcal{A}_{2,1,0}, \ldots, \mathcal{A}_{2,1,d}, \ldots, \mathcal{A}_{n,n,0}, \ldots \mathcal{A}_{n,n,d})^T,$$

that is vec$(\mathcal{A})$*) stacks the entry-wise coefficient vectors of each column on top of each other.*

We will find it convenient to define $x = x(\Delta\mathcal{A}, \mathcal{B})$ to be the combined vector of unknowns corresponding to $\Delta\mathcal{A}$ and $\mathcal{B}$. Let $\nabla^2_{xx}L$ denote the Hessian matrix of L with respect to x and J be the Jacobian of the residuals of the constraints, i.e. $J = \nabla_x M(\Delta\mathcal{A}, \mathcal{B})$. Necessary optimality conditions at a point (x^*, λ^*) (Bertsekas, 1999) are that

$$\nabla L = 0 \text{ and } \ker(J)^T \nabla^2_{xx} L \ker(J) \geq 0. \tag{4.5}$$

Sufficient conditions for optimality at the same point are that

$$\nabla L = 0 \text{ and } \ker(J)^T \nabla^2_{xx} L \ker(J) > 0. \tag{4.6}$$

These conditions are known as the second-order sufficiency conditions Bertsekas (1999). We note that (4.6) implies that minimal solutions are locally unique, and will fail to hold if minimal solutions are not locally unique. The idea is to show that (4.6) holds in the minimal embedding, which allows us to construct an algorithm with rapid local convergence.

4.3. The Jacobian

Definition 4.3. *The matrix $\psi(\widehat{\mathcal{b}})$ is an alternative form of $(\widehat{\mathcal{A}}+\widehat{\Delta\mathcal{A}})\widehat{\mathcal{b}} = 0$ that satisfies $\psi(\widehat{\mathcal{b}})\text{vec}(\mathcal{A}+\Delta\mathcal{A}) = 0$. That is, $\psi(\widehat{\mathcal{b}})$ satisfies*

$$\psi(\widehat{\mathcal{b}}) \cdot \text{vec}(\mathcal{A} + \Delta\mathcal{A}) = 0 \iff (\widehat{\mathcal{A}} + \widehat{\Delta\mathcal{A}})\widehat{\mathcal{b}} = 0.$$

We will adopt that notation that $\psi(\widehat{\mathcal{b}_i})$ corresponds to $\psi(\widehat{\mathcal{b}_i})\text{vec}(\widehat{\mathcal{A}_i} + \Delta\widehat{\mathcal{A}_i}) = 0$. Here we use the bi-linearity of (4.3) to write the same system using a matrix with entries from $\widehat{\mathcal{B}}$ instead of $\text{vec}(\mathcal{A} + \Delta\mathcal{A})$.

The closed-form expression for the Jacobian of the residuals (up to permutation) in (4.3) is given by (assuming $\mathcal{N}(\widehat{\mathcal{b}_j})\widehat{\mathcal{b}_j}$ is a quadratic function of $\widehat{\mathcal{b}_j}$)

$$J = \left(\begin{array}{c|ccccc} \psi(\widehat{\mathcal{b}_1}) & \widehat{\mathcal{A}_1} + \widehat{\Delta\mathcal{A}_1} & & & & \\ \psi(\widehat{\mathcal{b}_2}) & & \widehat{\mathcal{A}_2} + \widehat{\Delta\mathcal{A}_2} & & & \\ \vdots & & & \ddots & & \\ \psi(\widehat{\mathcal{b}_r}) & & & & \widehat{\mathcal{A}_r} + \widehat{\Delta\mathcal{A}_r} & \\ \hline 0 & 2\mathcal{N}(\widehat{\mathcal{b}_1})^T & & & & \\ 0 & & 2\mathcal{N}(\widehat{\mathcal{b}_2})^T & & & \\ \vdots & & & \ddots & & \\ 0 & & & & 2\mathcal{N}(\widehat{\mathcal{b}_r})^T & \end{array}\right). \tag{4.7}$$

Unlike the case of a single kernel vector in (Giesbrecht et al., 2017), J may be rank deficient since some equations corresponding to low (high) index entries may be redundant at the solution. The Lagrange multipliers will not be unique in this particular scenario and the rate of convergence may degrade if Newton's method is used. In the instance of $r = 1$ then we present the following result (Giesbrecht et al., 2017).

Theorem 4.4. *Suppose that $r = 1$ and $\widehat{\mathcal{b}_1}$ is minimally degree $\mathbb{R}$-embedded in $\widehat{\mathcal{A}_1}$, then J has full rank when* (4.5) *holds.*

Proof. We show that J has full row rank by contradiction. If this matrix was rank deficient, then one row is a linear combination of the others. This means that one of the equations in the constraints is trivial or the solution is not regular (see (Bertsekas, 1999, Section 3.1)). As we are only concerned about regular solutions, this contradicts the minimal $\mathbb{R}$-embedding. □

The corollary to this is that in the minimal embedding regularity conditions hold and it is straight forward to obtain rapid local convergence.

4.4. The Hessian

The Hessian matrix, $\nabla^2 L$ is straight forward to compute as

$$\nabla^2 L = \begin{pmatrix} \nabla^2_{xx} L & J^T \\ J & 0 \end{pmatrix}.$$

The following theorem shows that second-order sufficiency holds for the instance of $r = 1$. The case of $r > 1$ follows immediately by induction. This is in contrast to Theorem 4.4, which does not always hold for $r > 1$.

Theorem 4.5 (Second Order Sufficiency Holds)**.** *Suppose that $\widehat{\mathcal{A}} + \widehat{\Delta\mathcal{A}}$ has a minimally degree $\mathbb{R}$-embedded kernel vector $\widehat{\mathcal{b}}$, i.e. $r = 1$ in* (4.4)*, then at a minimal solution, the second order sufficiency condition* (4.6) *holds in the minimal embedding of $\widehat{\mathcal{b}}$.*

Proof. If $\|\Delta A\| = 0$ at the local minimizer (x^*, λ^*) then

$$\nabla^2_{xx} L(x^\star, \lambda^\star) = \begin{pmatrix} 2I & \\ & 0 \end{pmatrix} \text{ and } K = \ker \nabla^2_{xx} L(x^\star, \lambda^\star) = \operatorname{span} \begin{pmatrix} 0 & \\ & I \end{pmatrix}.$$

We have that for $y \in \operatorname{span}(K)$ such that $Jy = 0$ implies that $\widehat{\mathcal{A}} y = 0$ and $\mathcal{N}(\widehat{\mathcal{b}})^T y = 0$. It follows that $\ker \widehat{\mathcal{A}} = \operatorname{span}(\widehat{\mathcal{b}})$, thus we have $y = \widehat{\mathcal{b}}$ or $y = 0$ via the minimal $\mathbb{R}$-embedding, thus $y = 0$ as $\widehat{\mathcal{b}} \notin \operatorname{span}(K)$. Hence, second-order sufficiency holds, as $\ker J \cap K = 0$.

If $\|\Delta\mathcal{A}\| \neq 0$ then we have that

$$\nabla^2_{xx} L(x^\star, \lambda^\star) = \underbrace{\begin{pmatrix} 2I & 0 \\ 0 & 0 \end{pmatrix}}_{\mathcal{H}} + \underbrace{\begin{pmatrix} 0 & E^T \\ E & 0 \end{pmatrix}}_{\mathcal{E}}.$$

The matrix $\mathcal{E}$ is linear in λ, however the precise tensor decomposition is irrelevant to the proof. If E has full rank, then $\nabla^2_{xx} L$ has full rank and we are done, so suppose that E is rank deficient. If E is rank deficient, then one can eliminate a row of E and column of E^T without affecting $\mathcal{H}$ via symmetric row and column updates. We observe that $\ker(\mathcal{H} + \mathcal{E}) \subseteq \ker \mathcal{H}$ and the result follows. □

Corollary 4.6. *Suppose that $r > 1$ in* (4.4) *and $\mathcal{B}$ is minimally degree embedded, then second-order sufficiency* (4.6) *holds.*

Proof. The proof is almost the same as Theorem 4.5 and follows by induction on r since each block is decoupled. □

We now have all of the ingredients for an iterative method with rapid local convergence.

4.5. Iterative Post-Refinement

Newton's method for equality constrained minimization problems can be interpreted as solving the non-linear system of equations $\nabla L = 0$. Newton's method is based on the iterative update scheme

$$\begin{pmatrix} x^{k+1} \\ \lambda^{k+1} \end{pmatrix} = \begin{pmatrix} x^k + \Delta x^k \\ \lambda^k + \Delta \lambda^k \end{pmatrix} \text{ such that } \nabla^2 L \begin{pmatrix} \Delta x \\ \Delta \lambda \end{pmatrix} = -\nabla L. \tag{4.8}$$

If $r = 1$ then $\nabla^2 L$ has full rank and the iteration is well defined by matrix inversion. If $r > 1$ then we consider the quasi-Newton method defined as

$$\begin{pmatrix} x^{k+1} \\ \lambda^{k+1} \end{pmatrix} = \begin{pmatrix} x^k + \Delta x^k \\ \lambda^k + \Delta \lambda^k \end{pmatrix} \text{ such that } \begin{pmatrix} \nabla^2_{xx} L & J^T \\ J & -\mu_k I \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta \lambda \end{pmatrix} = -\nabla L \tag{4.9}$$

for a suitably chosen parameter μ_k. Taking $\mu_k = \|\nabla L(x^k, \lambda^k)\|_1$ one has provably quadratic convergence (Wright, 2005, Theorem 4.2) with x^k and λ^k chosen sufficiently close to the optimal solution.

Theorem 4.7. *The iteration* (4.9) *converges quadratically to* $(x^\star, \lambda^\star)$ *if* (x^0, λ^0) *are chosen sufficiently close to* $(x^\star, \lambda^\star)$.

We now have a method to compute a nearby rank deficient matrix polynomial with a rate of convergence that is quadratic, provided that the initial values of x are chosen to be sufficiently close to the optimal solution.

5. Implementation, Description and Comparison

In this section we discuss implementation details and demonstrate our implementation for computing the nearest rank deficient matrix polynomial. All algorithms are implemented in Maple 2016. All experiments are done using quad precision floating point arithmetic, with about 35 decimal digits of accuracy. We compare some degree one examples to the recent results of (Guglielmi et al., 2017).

To compute an approximate kernel vector, first we use the SVD to compute an approximate kernel of an $\mathbb{R}$-embedded (nearly) rank deficient matrix polynomial. Next we use structured orthogonal elimination RQ (LQ) decomposition to produce a minimally (degree) $\mathbb{R}$-embedded vector from the kernel. In the case of several kernel vectors we use a modified Gaussian elimination on an embedding of an approximate kernel obtained by the SVD and approximate GCD to find nearby approximate kernel vectors that are primitive.

5.1. Description of Algorithm

We now formally describe an algorithm for computing the nearest matrix polynomial of a prescribed rank. The algorithm has no global convergence guarantees, however a globally convergent (although not necessarily optimal) algorithm can be developed in a straight forward manner via augmenting our second order algorithm with a first order one, and removing content from kernel vectors if necessary.

The size of $\nabla^2 L$ is $O(r^2 n^4 d^2)$ and accordingly each iteration has a cost of $O(r^6 n^{12} d^6)$ flops using standard matrix multiplication, where r is the dimension of the kernel.

Algorithm 1 : `Iterative Kernel Post-Refinement`

Require:
- Full rank matrix polynomial $\mathcal{A} \in \mathbb{R}[t]^{n\times n}$
- (Approximately) Rank deficient matrix polynomial $\mathcal{C} \in \mathbb{R}[t]^{n]\times n}$
- Approximate kernel vectors $c_1, \ldots, c_r \in \mathbb{R}[t]^{n\times 1}$ of the desired degree/displacement structure
- Displacement structure matrix $\Delta\mathcal{A}$ to optimize over

Ensure:
- Singular matrix $\mathcal{A} + \Delta\mathcal{A}$ with $\mathcal{B} \subset \ker(\mathcal{A} + \Delta\mathcal{A})$ or an indication of failure.

1: $\mathbb{R}$-Embed $\mathcal{A}, \mathcal{C}, c_1, \ldots, c_r$ and $\Delta\mathcal{A}$.
2: Compute Lagrangian L from Section 4.2.
3: Initialize λ via linear least squares from $\nabla L|_x = 0$.
4: Compute $\begin{pmatrix} x + \Delta x \\ \lambda + \Delta\lambda \end{pmatrix}$ by solving (4.9) until $\left\| \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} \right\|_2$ is sufficiently small or divergence is detected.
5: Return the locally optimal $\Delta\mathcal{A}$ and $\mathcal{B}$ or an indication of failure.

5.2. Nearest Rank Deficient Linearly and Affinely Structured Matrix

In this section we consider Examples 2.10, 2.11 and 2.12 from Guglielmi et al. (2017), where we compare our results to real perturbations. Note that complex perturbations are a straightforward generalization of the theory presented here, and can be re-formulated as a problem over $\mathbb{R}$.

The technique of Guglielmi et al. (2017) poses computing a nearby rank-deficient linear matrix pencil by verifying that sufficiently many images of the matrix polynomial are singular, so that $\det(\mathcal{A} + \Delta\mathcal{A}) \equiv 0$. The problem is then posed as a solution to a system of Ordinary Differential Equations (ODE), assuming that certain genericity conditions on the eigenvalues of the solution hold[4]. They consider the instances of computing A_0 and A_1 with a common kernel vector, and the instance where A_0 and A_1 do not have a common kernel. Additionally, perturbations affecting only one of A_0 and A_1 are considered. We note that the solutions to the ODEs do not necessarily satisfy necessary optimality conditions (4.5), and accordingly will generally not be local minimizers.

5.2.1. Nearest Affinely Structured Examples I

Consider first the matrix polynomial

$$\mathcal{A} = \underbrace{\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}}_{A_1} t + \underbrace{\begin{pmatrix} 0 & 0.0400 & 0.8900 \\ 0.1500 & -0.0200 & 0 \\ 0.9200 & 0.1100 & 0.06600 \end{pmatrix}}_{A_0}$$

coming from Examples 2.10 and 2.12 of Guglielmi et al. (2017)

[4]Our algorithm and convergence theory does not explicitly rely on genericity assumptions or other properties of eigenvalues, however we do exploit generic properties in formulating initial guesses.

Example 5.1. *If we assume that A_1 is constant, then this is finding the (locally) nearest matrix polynomial with an affine structure since A_1 has non-zero fixed constants. First let's assume that zero entries are preserved, this is a linear structure on A_0.*

To compute an initial guess for $\mathcal{b}$ we use the SVD on $\widehat{\mathcal{A}}$ and extract a guess from the smallest singular vector. This gives us

$$\mathcal{b}_{init} = \begin{pmatrix} -0.41067t^3 + 0.50576t^2 - 0.26916t - 0.035720 \\ 0.38025t^2 - 0.51139t + 0.30674 \\ 0.027012t^2 - 0.028083t + 0.010715 \end{pmatrix}.$$

For an initial guess on $\mathcal{A}$ we take $\mathcal{A}_{init} = \mathcal{A}$. Note that we do not need an initial guess that is singular, it just needs to be "sufficiently close" to a singular matrix polynomial.

If we do not allow perturbations to zero-coefficients, that is, $A_0[1,1]$ and $A_0[2,3]$ may not be perturbed, then after five iterations of plain Newton's method (see (Giesbrecht et al., 2017)) we compute

$$\Delta A_0 \approx \begin{pmatrix} 0.0 & -0.094149 & -0.0057655 \\ -0.093311 & 0.026883 & 0.0 \\ 0.0057142 & -0.0016462 & -0.00010081 \end{pmatrix}$$

with perturbation $\|\Delta\mathcal{A}\|_F \approx 0.135507$.

A corresponding (approximate) kernel vector is

$$\mathcal{b} \approx \begin{pmatrix} 0.73073t + 0.082126 \\ -0.67644 \\ -0.041424 \end{pmatrix}.$$

Example 5.2. *If we allow perturbations to zero-coefficients in A_0 then after five rounds of plain Newton's method we compute*

$$\Delta A_0 \approx \begin{pmatrix} 0.0 & -0.094179 & -0.0057705 \\ -0.093280 & 0.026786 & 0.0016412 \\ 0.0057154 & -0.0016412 & -0.00010056 \end{pmatrix}$$

with perturbation $\|\Delta\mathcal{A}\|_F \approx 0.135497$, which is a marginal improvement over the previous example. A corresponding approximate kernel vector is

$$\mathcal{b} \approx \begin{pmatrix} 0.73073t + 0.082131 \\ -0.67644 \\ -0.041447 \end{pmatrix}.$$

Guglielmi et al. (2017) report an upper-bound on the distance to singularity allowing *complex perturbations*, that is $\Delta\mathcal{A} \in \mathbb{C}[t]^{n\times n}$ of $\|\Delta^{\mathbb{C}}\mathcal{A}\|_F \approx 0.1357$ in Example 2.10. In Example 2.12, Guglielmi et al. (2017) report an upper-bound on the distance to singularity allowing *real perturbations*, $\|\Delta^{\mathbb{R}}\mathcal{A}\|_F \approx 0.1366$. Although we only consider real perturbations, both bounds are improved. We conjecture that the complex bound can be improved further.

If we allow perturbations to A_0 and A_1, then this is some form of finding the nearest rank deficient matrix polynomial. The question is whether to allow degree or support preserving perturbations. Again, we will use the same initial guesses as the previous example.

Matrix degree preserving perturbations are of the form

$$\Delta^{deg}\mathcal{A} = \begin{pmatrix} tA_{1,1,1} + A_{1,1,0} & tA_{1,2,1} + A_{1,2,0} & tA_{1,3,1} + A_{1,3,0} \\ tA_{2,1,1} + A_{2,1,0} & tA_{2,2,1} + A_{2,2,0} & tA_{2,3,1} + A_{2,3,0} \\ tA_{3,1,1} + A_{3,1,0} & tA_{3,2,1} + A_{3,2,0} & tA_{3,3,1} + A_{3,3,0} \end{pmatrix},$$

where as support preserving perturbations are of the form

$$\Delta^{sup}\mathcal{A} = \begin{pmatrix} 0 & A_{1,2,0} & A_{1,3,0} \\ A_{2,1,0} & A_{2,2,0} & A_{2,3,1}t \\ A_{3,1,0} & tA_{3,2,1} + A_{3,2,0} & A_{3,3,0} \end{pmatrix}.$$

Example 5.3. *In the instance of degree preserving perturbations we compute after five iterations of Newton's method*

$$\Delta^{deg}\mathcal{A} \approx \begin{pmatrix} 0.0036502 & 0.0039174t - 0.066405 & 0.00011839t - 0.0020069 \\ -0.066897 & 0.058993t + 0.029807 & 0.0017829t + 0.00090082 \\ 0.0059893 & -0.0053098t - 0.0024133 & -0.00016047t - 0.000072934 \end{pmatrix}$$

with $\|\Delta^{deg}\mathcal{A}\| \approx 0.115585$.

A corresponding approximate kernel vector is

$$\mathcal{b} \approx \begin{pmatrix} -0.72941t - 0.080355 \\ 0.67903 \\ 0.020522 \end{pmatrix}.$$

Example 5.4. *In the instance of support preserving we compute after five iterations of Newton's method,*

$$\Delta^{sup}\mathcal{A} \approx \begin{pmatrix} 0.0 & -0.094311 & -0.0057928 \\ -0.092552 & 0.026973 & 0.0051028t \\ 0.0057434 & -0.0051554t - 0.0016739 & -0.00010281 \end{pmatrix}$$

with $\|\Delta^{sup}\mathcal{A}\| \approx 0.135313$. *A corresponding approximate kernel vector is*

$$\mathcal{b} \approx \begin{pmatrix} -0.72895t - 0.082339 \\ 0.67832 \\ 0.041664 \end{pmatrix}.$$

Guglielmi et al. (2017) report an upper-bound on the distance to singularity of $\|\Delta^{deg}\mathcal{A}\|_F \approx 0.1193$ in Example 2.12. This bound is larger than the one computed in Example 5.3.

5.3. Nearest Affinely Structured Examples II

Example 5.5. *Next we consider the the matrix polynomial* $\mathcal{A}$ *in Example 2.11 of (Guglielmi et al., 2017) defined as*

$$\mathcal{A} = \underbrace{\begin{pmatrix} -1.79 & 0.10 & -0.6 \\ 0.84 & -0.54 & 0.49 \\ -0.89 & 0.3 & 0.74 \end{pmatrix}}_{A_0} + \underbrace{\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}}_{A_1} t.$$

To compute an initial guess for we take $\mathcal{A}_{init} = A$ *and take*

$$b^{init} = \begin{pmatrix} -0.16001t^3 - 0.10520t^2 + 0.15811t + 0.11409 \\ 0.14980t^3 - 0.51289t^2 - 0.18616t + 0.54098 \\ 0.20801t^3 + 0.26337t^2 - 0.44619t - 0.027979 \end{pmatrix}.$$

b^{init} *is computed from the smallest singular vector of* $\widehat{\mathcal{A}}$.

We note that this initial guess does not attempt to find a nearby singular matrix polynomial for the initial guess, all that is needed is $\nabla L(x^{init}, \lambda^{init})$ *is reasonably small to obtain convergence.*

Using a globalized variant of Newton's method based on Levenberg-Marquardt we compute

$$\Delta\mathcal{A} = \begin{pmatrix} 0.047498t + 0.17772 & 0.44989t + 0.12420 & -0.091945t - 0.068210 \\ 0.20979t + 0.078872 & -0.094205t + 0.41583 & -0.037916t - 0.094081 \\ 0.082862t - 0.15413 & -0.58334t + 0.12940 & 0.081637t + 0.017208 \end{pmatrix},$$

with $\|\Delta\mathcal{A}\|_F \approx 0.949578$. *The corresponding approximate kernel vector is*

$$b = \begin{pmatrix} -0.29258t - 0.21491 \\ 0.044825t - 0.90281 \\ 0.068189t + 0.21562 \end{pmatrix}.$$

If we use the result of (Guglielmi et al., 2017) as the initial guess, then we compute

$$b^{init} = \begin{pmatrix} 0.16409t^2 + 0.25146t + 0.12362 \\ -4.5353 \times 10^{-14}t^2 + 0.23740t + 0.55516 \\ 1.2457 \times 10^{-13}t^2 - 0.48688t - 0.0060443 \end{pmatrix}.$$

We will assume the entries of b *are degree at most two.*

After five iterations of Newton's method we obtain

$$\Delta\mathcal{A} = \begin{pmatrix} 0.17257 & 0.12237t + 0.25225 & -0.46902t + 0.087147 \\ 0.21449 & 0.15210t + 0.31353 & -0.58296t + 0.10832 \\ -0.055963 & -0.039685t - 0.081803 & 0.15210t - 0.028261 \end{pmatrix},$$

with $\|\Delta\mathcal{A}\| \approx 0.94356416$.

The corresponding approximate kernel vector is

$$b = \begin{pmatrix} 0.18971t^2 + 0.29750t + 0.14667 \\ 0.27896t + 0.66186 \\ -0.58143t - 0.0079694 \end{pmatrix}.$$

The previously noted small quadratic terms were at roughly machine precision (the computation is done with 35 digits of precision) and truncated.

Guglielmi et al. (2017) obtain a result on this past example that produces an upper bound on the distance to singularity of 0.9438619. Their computation is accurate to seven decimal points, and accordingly our post-refinement has an improvement of about 0.000297. This is not surprising, since we solve the necessary conditions (4.5) directly with a reasonable initial guess.

5.4. Lower Rank Approximation of a 4×4 *Matrix*

In this following example we consider computing a lower-rank approximation to a given matrix polynomial. Consider the 4×4 matrix polynomial $\mathcal{A}$, defined as

$$\mathcal{A} = A_0 + A_1 t + A_2 t^2 + A_3 t^3, \text{ where}$$

$$A_0 = \begin{pmatrix} 0.09108776 & -0.05442464 & 0.3645006 & 0.01821543 \\ -0.1456436 & 0.03647524 & -0.07277662 & 0.07305016 \\ 0.05478714 & -0.05444916 & 0.4373220 & 0.05478385 \\ -0.1274211 & 0.09124859 & -0.6556615 & -0.05446850 \end{pmatrix},$$

$$A_1 = \begin{pmatrix} 0.09116729 & 0.00001797690 & 0.2550857 & 0.05475106 \\ 0.0001156514 & 0.00001659159 & 0.09108906 & -0.05447104 \\ 0.05470823 & 0.03662426 & 0.1276959 & 0.03650378 \\ 0.05472202 & -0.1091389 & 0.1458359 & -0.09090507 \end{pmatrix},$$

$$A_2 = \begin{pmatrix} 0.01833149 & 0.03661770 & 0.01824331 & 0.03660918 \\ 0.01837542 & -0.05442525 & 0.0 & 0.01832234 \\ 0.01841784 & 0.00003900436 & 0.0 & 0.01836515 \\ 0.01840752 & 0.00001508311 & 0.01839699 & 0.03659170 \end{pmatrix},$$

$$A_3 = \begin{pmatrix} 0.0 & 0.01837967 & 0.0 & 0.0 \\ 0.0 & 0.01843603 & 0.0 & 0.0 \\ 0.0 & 0.01829203 & 0.0 & 0.0 \\ 0.0 & 0.01842778 & 0.0 & 0.0 \end{pmatrix}.$$

Example 5.6. *We will consider a displacement structure on the kernel as well in this example, where higher-order zero terms are not perturbed from the initial guess. For the entries of* $\Delta\mathcal{A}$ *we preserve higher-order zero terms, and allow low-order terms to be perturbed. This is a linearly structured problem, on both the main variable* $\Delta\mathcal{A}$ *and the auxiliary kernel variable* $\mathcal{B}$*.*

To ensure the rank constraint holds, we will additionally assume that the kernel, $\widehat{\mathcal{B}}$ *is in a CREF (while* $\mathcal{B}$ *is obviously not) and the columns have unit norm. This normalization is (locally) equivalent to the ones discussed in Section 4.2. Having* $\widehat{\mathcal{B}}$ *in a CREF ensures that the two kernel vectors are locally linearly independent during the iteration. Of course perturbing both pivots to zero is possible (although this is sub-optimal). In such a scenario linear independence can no longer be guaranteed, and the iteration would need to be re-ininitialized.*

For the initial guess we use $\mathcal{A}^{init} = \mathcal{A}$ *and take* $\mathcal{B}^{init}$ *as*

$$\begin{pmatrix} 0.1954059t^2 & 0.0 \\ -0.2526800t - 0.7681472 & -0.06131396t^2 - 0.1839419t + 0.7357675 \\ -0.05727413t^2 - 0.01010720t - 0.1280246 & -0.06131396t^3 - 0.06131396t + 0.1226279 \\ 0.05727413t^2 + 0.4683004t + 0.2560491 & 0.06131396t^3 + 0.4905117t^2 - 0.3065698t - 0.2452558 \end{pmatrix}.$$

Using Algorithm 5.1 we compute after nine iterations

$$\Delta A_0 = \begin{pmatrix} 0.00003841866 & -0.0001970606 & -0.00002444167 & -0.000003273264 \\ 0.00001831140 & -0.00009026377 & 0.00002067189 & -0.0001255102 \\ -0.0001265513 & -0.0001595407 & 0.00003425737 & -0.00007523197 \\ -0.00007666528 & -0.0002773970 & 0.00004057408 & -0.0001720881 \end{pmatrix},$$

$$\Delta A_1 = \begin{pmatrix} 0.00001508776 & 0.00003166597 & 0.00004647888 & -0.0001142308 \\ -0.00005872595 & -0.00004487730 & 0.00004547421 & -0.0001483973 \\ 0.00002056901 & -0.0001596527 & -0.000006413632 & -0.00006541721 \\ -0.00003695701 & -0.0001773889 & 0.00004119722 & -0.0002159825 \end{pmatrix},$$

$$\Delta A_2 = \begin{pmatrix} -0.00003352295 & -0.0001190577 & 0.00005687700 & -0.0001783770 \\ 0.00001768442 & -0.0001467423 & 0.0 & -0.00008587235 \\ -0.00006506345 & 0.00005243135 & 0.0 & -0.0001686619 \\ -0.0001471227 & -0.0001295490 & -0.00001105246 & -0.0001124559 \end{pmatrix},$$

$$\Delta A_3 = \begin{pmatrix} 0.0 & -0.0001025690 & 0.0 & 0.0 \\ 0.0 & -0.0001315095 & 0.0 & 0.0 \\ 0.0 & -0.00002763942 & 0.0 & 0.0 \\ 0.0 & -0.0001877673 & 0.0 & 0.0 \end{pmatrix},$$

with $\|\Delta\mathcal{A}\| \approx 0.0007844$.

An approximate kernel, $\mathcal{B}$ *is given by*

$$\begin{pmatrix} 0.1955493t^2 + 0.0006874986t - 0.001013023 & 0.0 \\ -0.2542383t - 0.7686061 & -0.06128819t^2 - 0.1818298t + 0.7368313 \\ -0.05698735t^2 - 0.01004111t - 0.1276311 & -0.06125293t^3 - 0.0002486115t^2 - 0.06112324t + 0.1226783 \\ 0.05795811t^2 + 0.4677475t + 0.2541290 & 0.06151690t^3 + 0.4894569t^2 - 0.3069667t - 0.2452396 \end{pmatrix}.$$

A natural question is what happens if we change the displacement structure on the kernel? To investigate this behavior, we consider an equivalent representation of the previously used kernel, except that $\mathcal{B}$ is in a CREF directly.

Example 5.7. *If we change the kernel* $\mathcal{B}^{init}$ *to be*

$$\begin{pmatrix} 0.1581139t^3 + 0.1581139t - 0.3162278 & 0.03965258t^3 + 0.3172206t^2 - 0.1982629t - 0.1586103 \\ -0.1581139t^2 - 0.4743417t - 0.6324556 & -0.03965258t^2 - 0.4361784t - 0.7930516 \\ 0.0 & 0.07930516t - 0.07930516 \\ 0.3162278t - 0.3162278 & 0.0 \end{pmatrix},$$

used in the initialization of the previous example, then we compute a perturbation with $\|\Delta\mathcal{A}\| \approx 0.0008408$.

In either case, we obtain comparable answers that are a reasonable lower-rank approximation, and can likely be improved by relaxing restrictions on the displacement structure on $\mathcal{B}$ or $\widehat{\mathcal{B}}$. It is important to note that relaxing the degree bounds to be $(n-r)d$ in general on all non-zero entries (where entries are zero if they are in the same row as a CREF pivot) will likely lead to a better approximation, however one may lose quadratic convergence if doing so, since iterates may no longer have primitive kernel vectors, and (4.6) will no longer hold. As discussed in Section 4, it is generally difficult to determine the CREF pivots of the kernel unless the initial guess is very accurate.

The structure of the kernel is an important consideration when deciding upon an initial guess. It is preferable to restrict fewer coefficients, however the iteration requires a better initialization due to the increased number of possible descent directions. In such scenarios for maximum flexibility, a globalized variant of Newton's method is required. Like-wise, the structure for $\Delta\mathcal{A}$ is also an important choice. Restricting which terms can be changed has a large influence on the (approximate) distance to singularity (of prescribed kernel dimension).

Another way to approach the lower-rank approximation problem is to use alternating projections or alternating directions of descent (since the objective is bi-linear with bi-linear constraints, it is convex in each argument) on the rank factorization in Section 3. Since solutions in one coordinate, $\Delta\mathcal{A}$ are isolated, one can expect linear convergence with a reasonable algorithm. The lack-of normalization required overcomes the difficulty of choosing a suitable kernel displacement structure, however convergence would be linear at best and determining the dimensions of U and V is another problem to be discussed. It is also worth noting that Algorithm 5.1 requires more computational resources per iteration as r increases, however a rank factorization requires fewer computational resources per iteration as r increases.

6. Conclusions and Future Work

We have shown that finding lower-rank approximations of matrix polynomials can be established as a numerically well-posed problem and is amenable to first and second order optimization methods. The existence and isolation of solutions is established along with an algorithm exploiting affine structures to obtain locally quadratic convergence under mild normalization assumptions.

Along with considering the lower-rank approximation of matrix polynomials, we present a generalization of the theory to matrix polynomials with an arbitrary affine structure. We provide examples of how the structure of permissible perturbations and prescribed kernel structure impacts the distance to solutions.

We also regard this current paper as a first step towards a formally robust approach to nonlinear matrix polynomials, in the spirit of recent work with symbolic-numeric algorithms for polynomials. Problems such as approximate matrix polynomial division, GCRD and factorization all have applications which can benefit from these modern tools.

References

Abatzoglou, T., Mendel, J., Harada, G., 1991. The constrained total least squares technique and its applications to harmonic superresolution. IEEE Transactions on Signal Processing 39 (5), 1070–1087.

Absil, P.-A., Mahony, R., Sepulchre, R., 2009. Optimization algorithms on matrix manifolds. Princeton University Press.

Beckermann, B., Labahn, G., 1998a. A fast and numerically stable euclidean-like algorithm for detecting relatively prime numerical polynomials. Journal of Symbolic Computation 26 (6), 691–714.

Beckermann, B., Labahn, G., 1998b. When are two numerical polynomials relatively prime? Journal of Symbolic Computation 26, 677–689.

Beckermann, B., Labahn, G., Villard, G., 2006. Normal forms for general polynomial matrices. Journal of Symbolic Computation 41 (6), 708–737.

Bertsekas, D., 1999. Nonlinear programming. Athena Scientific.

Botting, B., Giesbrecht, M., May, J., 2005. Using the Riemannian SVD for problems in approximate algebra. In: Proc. Workshop on Symbolic-Numeric Computation. pp. 209–219.

Braatz, R. P., Young, P. M., Doyle, J. C., Morari, M., 1994. Computational complexity of μ calculation. IEEE Transactions on Automatic Control 39 (5), 1000–1002.

Byers, R., He, C., Mehrmann, V., 1998. Where is the nearest non-regular pencil? Linear Algebra and its Applications 285 (1), 81–105.

Byers, R., Nichols, N., 1993. On the stability radius of a generalized state-space system. Linear Algebra and its Applications 188, 113–134.

Cadzow, J., 1988. Signal enhancement – a composite property mapping algorithm. IEEE Transactions on Acoustics, Speech, and Signal Processing 36 (1), 49–62.

Corless, R. M., Gianni, P., Trager, B., Watt, S., 1995. The singular value decomposition for polynomial systems. In: Proc. International Symposium on Symbolic and Algebraic Computation (ISSAC). pp. 96–103.

De Moor, B., 1993. Structured total least squares and L_2 approximation problems. Linear Algebra and its Applications 188, 163–205.

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.

De Moor, B., 1995. The Riemannian singular value decomposition. In: Signal Processing, III: Algorithms, Architectures and Applications. Elsevier, pp. 61–78.

Eckart, G., Young, G., 1936. The approximation of one matrix by another of lower rank. Psychometrica 1, 211–218.

Giesbrecht, M., Haraldson, J., 2014. Computing GCRDs of approximate differential polynomials. In: Proc. Workshop on Symbolic-Numeric Computing (SNC'14). pp. 78–87.

Giesbrecht, M., Haraldson, J., G., L., 2017. Computing the nearest rank-deficient matrix polynomial. In: Proc. International Symposium on Symbolic and Algebraic Computation (ISSAC). ACM Press, pp. 181–188.

Giesbrecht, M., Haraldson, J., Kaltofen, E., 2016. Computing approximate greatest common right divisors of differential polynomials, submitted for publication, 41 pages.

Gohberg, I., Lancaster, P., Rodman, L., 2009. Matrix Polynomials. SIAM.

Golub, G., Pereyra, V., 1973. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. SIAM Journal on Numerical Analysis 10 (2), 413–432.

Golub, G., Pereyra, V., 2003. Separable nonlinear least squares: the variable projection method and its applications. Inverse problems 19 (2), R1.

Golub, G., Van Loan, C., 2012. Matrix Computations. Vol. 3. JHU Press.

Guglielmi, N., Lubich, C., Mehrmann, V., 2017. On the nearest singular matrix pencil. SIAM. Journal on Matrix Analysis and Applications (to appear).

Henrion, D., Lasserre, J.-B., 2006. Convergent relaxations of polynomial matrix inequalities and static output feedback. IEEE Transactions on Automatic Control 51 (2), 192–202.

Kaltofen, E., May, J. P., Yang, Z., Zhi, L., 2008. Approximate factorization of multivariate polynomials using singular value decomposition. Journal of Symbolic Computation 43 (5), 359–376.

Kaltofen, E., Storjohann, A., 2015. The complexity of computational problems in exact linear algebra. In: Encyclopedia of Applied and Computational Mathematics. Springer, pp. 227–233.

Kaltofen, E., Yang, Z., Zhi, L., 2005. Structured low rank approximation of a sylvester matrix. In: Proc. Int. Workshop on Symbolic-Numeric Computation (SNC 2005). pp. 188–201.

Kaltofen, E., Yang, Z., Zhi, L., 2006. Approximate greatest common divisors of several polynomials with linearly constrained coefficients and singular polynomials. In: Proc. International Symposium on Symbolic and Algebraic Computation (ISSAC). ACM Press, pp. 169–176.

Kressner, D., Voigt, M., 2015. Distance problems for linear dynamical systems. In: Numerical Algebra, Matrix Theory, Differential-Algebraic Equations and Control Theory. Springer, pp. 559–583.

Lasserre, J.-B., 2001. Global optimization with polynomials and the problem of moments. SIAM Journal on Optimization 11 (3), 796–817.

Lawrence, P. W., Corless, R. M., 2015. Backward error of polynomial eigenvalue problems solved by linearization of lagrange interpolants. SIAM J. Matrix Analysis 36 (4), 1425–1442.

Li, B., Yang, Z., Zhi, L., 2005. Fast low rank approximation of a Sylvester matrix by structure total least norm. J. Japan Soc. Symbolic and Algebraic Computation 11, 165–174.

Markovsky, I., 2008. Structured low-rank approximation and its applications. Automatica 44 (4), 891–909.

Markovsky, I., 2011. Low rank approximation: algorithms, implementation, applications. Springer Science & Business Media.

Nie, J., 2014. Optimality conditions and finite convergence of lasserre's sos hierarchy. Mathematical Programming 146 (1-2), 97–121.

Poljak, S., Rohn, J., 1993. Checking robust nonsingularity is np-hard. Mathematics of Control, Signals, and Systems (MCSS) 6 (1), 1–9.

Rosen, B., Park, H., Glick, J., 1996. Total least norm formulation and solution for structured problems. SIAM Journal on Matrix Analysis and Applications 17 (1), 110–126.

Rosen, B., Park, H., Glick, J., 1998. Structured total least norm for nonlinear problems. SIAM Journal on Matrix Analysis and Applications 20 (1), 14–30.

Schost, É., Spaenlehauer, P.-J., 2016. A quadratically convergent algorithm for structured low-rank approximation. Foundations of Computational Mathematics 16 (2), 457–492.

Storjohann, A., Villard, G., 2005. Computing the rank and a small nullspace basis of a polynomial matrix. In: Proc. International Symposium on Symbolic and Algebraic Computation (ISSAC). ACM Press, pp. 309–316.

Wright, S., 2005. An algorithm for degenerate nonlinear programming with rapid local convergence. SIAM Journal on Optimization 15 (3), 673–696.

Zhi, L., 2007. Numerical optimization in hybrid symbolic-numeric computation. In: Proc. 2007 International Workshop on Symbolic-Numeric Computation. pp. 33–35.