
ELSEVIER Linear Algebra and its Applications 343–344 (2002) 171–193

LINEAR ALGEBRA AND ITS APPLICATIONS

www.elsevier.com/locate/laa

Inversion error, condition number, and approximate inverses of uncertain matrices

Laurent El Ghaoui

Department of Electrical Engineering and Computer Science, University of California at Berkeley, Berkeley, CA 94720, USA

Received 14 July 2000; accepted 23 January 2001

Submitted by A.H. Sayed

Abstract

The classical condition number is a very rough measure of the effect of perturbations on the inverse of a square matrix. First, it assumes that the perturbation is infinitesimally small. Second, it does not take into account the perturbation structure (e.g., Vandermonde). Similarly, the classical notion of the inverse of a matrix neglects the possibility of large, structured perturbations. We define a new quantity, the structured maximal inversion error, that takes into account both structure and non-necessarily small perturbation size. When the perturbation is infinitesimal, we obtain a "structured condition number". We introduce the notion of approximate inverse, as a matrix that best approximates the inverse of a matrix with structured perturbations, when the perturbation varies in a given range.

For a wide class of perturbation structures, we show how to use (convex) semidefinite programming to compute bounds on the structured maximal inversion error and structured condition number, and compute an approximate inverse. The results are exact when the perturbation is "unstructured"—we then obtain an analytic expression for the approximate inverse. When the perturbation is unstructured and additive, we recover the classical condition number; the approximate inverse is the operator related to the Total Least Squares (orthogonal regression) problem.

Keywords: Structured matrix; Condition number; Linear fractional representation; Semidefinite programming; Vandermonde system; Total least squares

Notation

For a matrix X, $\|X\|$ denotes the largest singular value. If X is square, $X \geqslant 0$ (resp. $X > 0$) means X is symmetric, and positive semidefinite (resp. definite). For

E-mail address: elghaoui@eecs.berkeley.edu (L. El Ghaoui).

0024-3795/02/$ - see front matter
PII: S0024-3795(01)00273-7

a vector x, $\max_i |x_i|$ is denoted by $\|x\|_\infty$. The notation I_p (resp. $0_{p\times q}$) denotes the $p \times p$ identity (resp. $p \times q$ zero) matrix; sometimes the subscript is omitted when it can be inferred from context. To a given linear set $\mathit{\Delta} \subseteq \mathbb{R}^{p\times q}$, we associate the linear subspace $\mathscr{B}(\mathit{\Delta})$, defined by

$$\mathscr{B}(\mathit{\Delta}) = \{(S, T, G) \mid S\Delta = \Delta T,\ G\Delta = -\Delta^{\mathrm{T}} G^{\mathrm{T}} \text{ for every } \Delta \in \mathit{\Delta}\}. \tag{0.1}$$

1. Introduction

1.1. Motivations

Let $A \in \mathbb{R}^{n\times n}$, $\det A \neq 0$. We consider the problem of measuring, and reducing, the effect of errors when computing the inverse of A.

When the error on A, ΔA, is infinitesimally small, and otherwise arbitrary, a classical result (see e.g. , [6]) states that

$$\frac{\|(A + \Delta A)^{-1} - A^{-1}\|}{\|A^{-1}\|} \leqslant \kappa(A)\frac{\|\Delta A\|}{\|A\|}, \tag{1.1}$$

where $\kappa(A) = \|A\| \cdot \|A^{-1}\|$. Thus, the classical condition number $\kappa(A)$ is a measure of (relative) errors in the inverse of A when the latter is perturbed by an arbitrary, infinitesimally small matrix. (Note that the "absolute" condition number is then $\|A^{-1}\|^2$.)

The classical condition number is a very rough measure of the effect of errors when inverting A. First, the condition number above assumes that each coefficient in A is independently perturbed, which is often unrealistic. For instance, if A has a Toeplitz or Vandermonde structure, the perturbation matrix ΔA inherits the same structure. Therefore, the "structured condition number" is expected to be less than $\kappa(A)$ [13]. Second, the error bound (1.1) is only valid for (infinitesimally) small perturbations.

The classical definition of the inverse of a matrix also neglects the possibility of large perturbations. Consider the scalar equation $\mathbf{a}x = 1$, where $\mathbf{a}$ is unknown-but-bounded, say, $\mathbf{a} \in \mathscr{I} = [a - \rho \quad a + \rho]$, where ρ $(0 < \rho < |a|)$ is given. The possible values of the solution lie in the interval $\mathscr{J} = [(a - \rho)^{-1} \quad (a + \rho)^{-1}]$. Without more information about the "distribution" of $\mathbf{a}$ in the interval $\mathscr{I}$, the "best" value of the inverse is not a^{-1} (the classical inverse). A more accurate value is the *center* of the interval $\mathscr{J}$, that is, $a/(a^2 - \rho^2)$.

Perturbation structure is also neglected in the classical definition of a (matrix) inverse. Consider again a scalar equation $\mathbf{a}x = 1$, where $\mathbf{a} = \mathbf{c}^2$, and the "Cholesky factor" $\mathbf{c}$ is unknown-but-bounded (say, $\mathbf{c} \in \mathscr{I} = [c - \rho \quad c + \rho]$). As before, we may define an "approximate inverse" as the center of the set of possible values of $\mathbf{c}^{-2}$, which is $(a + \rho^2)/(a - \rho^2)^2$. Note that this value is in general different from its "unstructured" counterpart.

1.2. Framework

The above remarks call for a precise study of the effect of non-necessarily small, possibly nonlinear, structured perturbations, on the inverse of A. For this we introduce a very general model for the perturbation structure. We assume that the perturbation is a $p \times q$ matrix Δ that is restricted to a given linear subspace $\boldsymbol{\Delta} \subseteq \mathbb{R}^{p\times q}$. We then assume that the perturbed value of A can be written in the "linear-fractional representation" (LFR)

$$\mathbf{A}(\Delta) = A + L\Delta\,(I - D\Delta)^{-1}\,R, \tag{1.2}$$

where A is the (square, invertible) "nominal" value, and L, R, D are given matrices of appropriate size (the above expression is not always well defined; we return to this issue soon). The norm used to measure the perturbation size is the largest singular value norm, $\|\Delta\|$. For a given $\rho \geqslant 0$, we define the *perturbation* set by

$$\boldsymbol{\Delta}_\rho = \{\Delta \in \boldsymbol{\Delta} \mid \|\Delta\| \leqslant \rho\}\ .$$

The above model seems very specialized, but it can be used for a very wide variety of perturbation structures (see Section 2). In particular, our framework includes the case when parameters perturb each coefficient of the data matrices linearly, and in addition, the parameters are bounded componentwise.

Our subject is the study of the following notions.

The *invertibility radius*, denoted $\rho^{\text{inv}}(\mathbf{A}, \boldsymbol{\Delta})$, is the largest value of ρ such that $\mathbf{A}(\Delta)$ is well-posed (in the sense that $\det(I - D\Delta) \neq 0$) and invertible for every $\Delta \in \boldsymbol{\Delta}_\rho$.

For $0 < \rho < \rho^{\text{inv}}(\mathbf{A}, \boldsymbol{\Delta})$, we define the *structured maximal inversion error* as

$$\lambda(\mathbf{A}, \boldsymbol{\Delta}, \rho) = \frac{1}{\rho} \max \left\{ \|\mathbf{A}(\Delta)^{-1} - A^{-1}\| \ :\ \Delta \in \boldsymbol{\Delta}_\rho \right\}. \tag{1.3}$$

We define the *structured absolute condition number* by

$$\kappa(\mathbf{A}, \boldsymbol{\Delta}) = \limsup_{\rho \to 0} \lambda(\mathbf{A}, \boldsymbol{\Delta}, \rho).$$

Finally, we say that X is an *approximate inverse over* $\boldsymbol{\Delta}_\rho$ for the structured matrix $\mathbf{A}$ if it minimizes the *maximal inversion error at* X, defined as

$$\lambda(\mathbf{A}, \boldsymbol{\Delta}, \rho, X) = \frac{1}{\rho} \max_{\Delta} \left\{ \|\mathbf{A}(\Delta)^{-1} - X\| \ :\ \Delta \in \boldsymbol{\Delta}_\rho \right\}. \tag{1.4}$$

The approximate inverse is defined as the center of a ball that contains the possible values of the inverse $\mathbf{A}(\Delta)^{-1}$, when Δ varies over the perturbation set $\boldsymbol{\Delta}_\rho$. In this sense, the approximate inverse generalizes the scalar case mentioned in Section 1.1.

The problems addressed in this paper are in general NP-complete. Our purpose is to compute bounds for these problems, via *semidefinite programming*. A (generalized) semidefinite program (SDP) is a problem of the form

$$\text{minimize } \lambda \quad \text{subject to } \lambda B(x) - A(x) \geqslant 0, \quad B(x) \geqslant 0, \quad C(x) \geqslant 0, \quad (1.5)$$

where $A(\cdot)$, $B(\cdot)$ and $C(\cdot)$ are affine functions taking values in the space of symmetric matrices, and $x \in \mathbb{R}^m$ is the variable. SDPs are (quasi-) convex optimization problems and can be solved in polynomial-time with e.g. , primal–dual interior-point methods [3,18,19,27]. Our approach thus leads to polynomial-time algorithms for computing bounds for our problems. In some cases, we obtain analytic expressions involving no iterative algorithms.

In this paper, we compute quantities associated to the matrix-valued function **A** via the LFR (1.2). Thus, we make no distinction between the matrix function **A** and its LFR (1.2), although in principle different LFRs of the same matrix-valued function **A** might give different numbers. It turns out, however, that the quantities we compute are independent, in some sense, of the LFR chosen to describe **A**. We make this sense precise in Appendix A.

1.3. Previous work

A complete bibliography on structured perturbations in linear algebra is clearly out of scope here. Many chapters of the excellent book by Higham [14] are relevant, especially the parts on error bounds for linear systems (pp. 143–145), condition number estimation (Chapter 14) and automatic error analysis (Chapter 24). The present paper is also related to interval arithmetic computations, which is a large field of study, since its introduction by Moore [15,16]. We briefly comment on this connection in Section 8.

The invertibility radius is related to the notion of nonsingularity radius (or distance to the nearest singular matrix). Most authors concentrated on the case when the perturbation enters affinely in $\mathbf{A}(\Delta)$. Even in this case, computing this quantity is NP-hard, see [17,20]. Demmel [6] and Rump [24] discuss bounds for the nonsingularity radius in this case. The bound proposed here is a variant of that given by Fan et al. in [9].

The maximal inversion error is closely related to systems of linear interval equations (which are covered by LFR models). Exact (NP-hard) bounds on (interval) solutions to such systems are discussed by Rohn in [21–23]. Alternative norms for measuring the error can be used, as pointed out by Hagher [12].

The structured condition number problem is addressed by Bartels and Higham [2] and by Gohberg and Koltracht [11]. The approach is based on the differentiation of a mapping describing the perturbation structure, which gives information on the effect of infinitesimal perturbations.

Matrix structures are described by a variety of tools. The displacement-rank model is one, see [5,10]. The LFR models used here are classical in robust control (see e.g. [4]). These models are used in the context of least squares problems with uncertain data by the authors in [7]. The results presented here can be viewed as extensions of the results proposed in [7].

2. Examples of LFR models

Before addressing the problems defined in Section 1, we first illustrate how the LFR model can be used in a variety of situations.

2.1. Additive perturbations with norm bound

The additive perturbations case is when

$$\mathbf{A}(\Delta) = A + \Delta,$$

and Δ is norm-bounded and otherwise arbitrary. This structure, the simplest of all we consider, corresponds to the matrix defined in (1.2), with $L = I_n$, $R = I_n$, $D = 0_n$ and $\boldsymbol{\Delta} = \mathbb{R}^{n\times n}$. The set $\mathscr{B}(\boldsymbol{\Delta})$ associated with $\boldsymbol{\Delta}$, as defined in (0.1), takes the form

$$\mathscr{B}(\boldsymbol{\Delta}) = \{(S, T, G) \mid S = T = \tau I_n,\ G = 0_n,\ \tau \in \mathbb{R}\}\ . \tag{2.1}$$

The additive model will be useful to recover classical results such as the standard condition number.

2.2. Unstructured perturbations with norm bound

The case when $\boldsymbol{\Delta} = \mathbb{R}^{p\times q}$ is referred to as the "unstructured perturbations case". This is a generalization of the additive model, that is useful to model perturbations that occur e.g. in only some columns (or rows) of A, but are otherwise arbitrary. The set $\mathscr{B}(\boldsymbol{\Delta})$ associated with $\boldsymbol{\Delta} = \mathbb{R}^{p\times q}$, as defined in (0.1), takes the form

$$\mathscr{B}(\boldsymbol{\Delta}) = \left\{(S, T, G) \;\middle|\; S = \tau I_p,\ T = \tau I_q,\ G = 0,\ \tau \in \mathbb{R}\right\}\ . \tag{2.2}$$

Consider for example the case when $\mathbf{A}$ can be partitioned as

$$\mathbf{A}(\Delta) = \begin{bmatrix} A_1 + \Delta \\ A_2 \end{bmatrix},$$

where $A_1 \in \mathbb{R}^{(n-r)\times n}$, $A_2 \in \mathbb{R}^{r\times n}$ are given, the perturbation matrix Δ is norm-bounded, and otherwise arbitrary. This case happens when we assume additive perturbations on some rows of A only. We may model this perturbation structure by (1.2), with

$$A = \left[\frac{A_1}{A_2}\right], \quad L = \left[\frac{I_{n-r}}{0_{r\times(n-r)}}\right], \quad R = I_n, \quad D = 0.$$

2.3. Affine perturbation with componentwise bound

As said before, our framework includes the case when parameters perturb each coefficient of the data matrices linearly, and in addition, the parameters are

bounded componentwise. Consider a matrix-valued functions $\mathbf{A}(\delta)$ that is affine in $\delta \in \mathbb{R}^m$:

$$\mathbf{A}(\delta) \stackrel{\Delta}{=} A_0 + \sum_{i=1}^{m} \delta_i A_i,$$

where $A_0, \ldots, A_p \in \mathbb{R}^{n\times n}$, are given. We can write $\mathbf{A}(\delta)$ in the LFR format, as follows.

For every i, $i = 1, \ldots, m$, decompose A_i as $A_i = L_i R_i$, with $L_i \in \mathbb{R}^{n\times r_i}$, $R_i \in \mathbb{R}^{r_i\times n}$, where $r_i = \mathbf{Rank}(A_i)$. With

$$L = [L_1 \ \ldots \ L_m], \quad R = \left[R_1^{\mathrm{T}} \ \ldots \ R_m^{\mathrm{T}}\right]^{\mathrm{T}},$$

we have

$$\mathbf{A}(\delta) = A + L\Delta R,$$

where $\Delta = \mathbf{diag}\left(\delta_1 I_{r_1}, \ldots, \delta_m I_{r_m}\right)$. Componentwise bounds of the form $\|\delta\|_\infty \leqslant \rho$ can be written $\|\Delta\| \leqslant \rho$.

In this case, the set $\boldsymbol{\Delta}$ we work with is of the form

$$\boldsymbol{\Delta} = \left\{\Delta = \mathbf{diag}\left(\delta_1 I_{r_1}, \ldots, \delta_m I_{r_m}\right),\ \delta \in \mathbb{R}^m\right\}, \tag{2.3}$$

and the subspace $\mathscr{B}(\boldsymbol{\Delta})$ associated to $\boldsymbol{\Delta}$ is

$$\mathscr{B}(\boldsymbol{\Delta}) = \left\{ (S, T, G) \;\middle|\; \begin{array}{l} S = T = \mathbf{diag}(S_1, \ldots, S_m),\ S_i \in \mathbb{R}^{r_i\times r_i}, \\ G = \mathbf{diag}(G_1, \ldots, G_m), \\ \quad G_i \in \mathbb{R}^{r_i\times r_i},\ G = -G^{\mathrm{T}} \end{array} \right\}. \tag{2.4}$$

2.4. Rational perturbations with componentwise bound

Our framework includes the case when parameters perturb each coefficient of the data matrices in a (polynomial or) rational manner. This is thanks to the representation lemma given below.

Lemma 2.1. *For any rational matrix function* $\mathbf{M} : \mathbb{R}^m \to \mathbb{R}^{n\times c}$, *with no singularities at the origin, there exist nonnegative integers* $r_1, \ldots, r_m$, *and matrices* $M \in \mathbb{R}^{n\times c}$, $L \in \mathbb{R}^{n\times N}$, $R \in \mathbb{R}^{N\times c}$, $D \in \mathbb{R}^{N\times N}$, *with* $N = r_1 + \cdots + r_m$, *such that* $\mathbf{M}$ *has the following linear-fractional representation* (*LFR*):

$$\begin{array}{l} \boldsymbol{M}(\delta) = M + L\Delta\,(I - D\Delta)^{-1} R, \\ \quad \textit{where}\ \ \Delta = \mathbf{diag}\left(\delta_1 I_{r_1}, \ldots, \delta_m I_{r_m}\right), \end{array} \tag{2.5}$$

valid for every δ *such that* $\det(I - D\Delta) \neq 0$.

A linear-fractional representation (LFR) is thus a matrix-based way to describe a multivariable rational matrix-valued function. It is a generalization, to the multivariable case, of the well-known state-space representation of transfer functions.

In [28], a constructive proof of the above result is given. The proof is based on a simple idea: first devise LFRs for simple (e.g. linear) functions, then use combination rules (such as multiplication, addition, etc.), to devise LFRs for arbitrary rational functions. Note that such a construction of the LFR can be done in polynomial-time.

The implication of the lemma for the study of structured condition numbers is far-reaching. If we consider a matrix-valued function $\mathbf{A}(\delta)$ that is (arbitrary) rational functions of a parameter vector $\delta \in \mathbb{R}^m$, it is possible to form the LFR of the function $\mathbf{A}(\delta)$, as in done in lemma 2.1. In this case, the set $\boldsymbol{\Delta}$ we work with is of the form (2.3) where r is the integer m-vector appearing in the LFR (2.5). The subspace $\mathscr{B}(\boldsymbol{\Delta})$ associated to $\boldsymbol{\Delta}$ is given by (2.4). Finally, note that the bound $\|\Delta\| \leqslant \rho$ is equivalent to componentwise bounds on the perturbation vector δ: $\|\delta\|_\infty \leqslant \rho$.

As an example, consider the square Vandermonde matrix

$$\mathbf{A}(a) = \begin{bmatrix} 1 & a_1 & \cdots & a_1^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & a_n & \cdots & a_n^{n-1} \end{bmatrix}, \tag{2.6}$$

where $a = (a_1, \ldots, a_n)^{\mathrm{T}} \in \mathbb{R}^n$ is a given vector. We assume that a is subject to componentwise, unstructured perturbation. That is, the perturbed value of a is $a + \delta$, where $\|\delta\|_\infty \leqslant \rho$. The perturbed matrix $\mathbf{A}(a + \delta)$ can be expressed with the LFR (1.2), where $A = \mathbf{A}(a)$, and

$$\begin{aligned} &L = \mathbf{diag}_{i=1}^n \begin{bmatrix} 1 & a_i & \cdots & a_i^{n-2} \end{bmatrix}, \quad R = \begin{bmatrix} R_1 \\ \vdots \\ R_n \end{bmatrix}, \\ &D = \mathbf{diag}_{i=1}^n D_i, \quad \Delta = \mathbf{diag}_{i=1}^n \delta_i I_{n-1}, \end{aligned} \tag{2.7}$$

and, for each i, $i = 1, \ldots, n$,

$$\begin{aligned} R_i &= \begin{bmatrix} 0 & 1 & a_i & \cdots & a_i^{n-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & a_i \\ 0 & \ldots & \ldots & 0 & 1 \end{bmatrix} \in \mathbb{R}^{(n-1)\times n}, \\ D_i &= \begin{bmatrix} 0 & 1 & a_i & \cdots & a_i^{n-3} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & a_i \\ \vdots & & & \ddots & 1 \\ 0 & \cdots & \cdots & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{(n-1)\times(n-1)}. \end{aligned} \tag{2.8}$$

3. Invertibility radius

3.1. Well-posedness lemma

We are given a linear-fractional function of a matrix variable $\Delta \in \boldsymbol{\Delta}$

$$\mathbf{M}(\Delta) = M + L\Delta\,(I - D\Delta)^{-1}\,R. \tag{3.1}$$

We say that the LFR (3.1) is well-posed over $\boldsymbol{\Delta}_\rho$ if

$$\det(I - D\Delta) \neq 0 \quad \text{for every } \Delta \in \boldsymbol{\Delta}_\rho.$$

The *well-posedness radius* of $\boldsymbol{M}$ is the largest ρ such that the LFR (3.1) is well-posed over $\boldsymbol{\Delta}_\rho$. The following lemma is a variant of a result first given in [9].

Lemma 3.1. *The LFR of $\boldsymbol{M}$ is well-posed over $\boldsymbol{\Delta}_\rho$ if there exists a triple (S, T, G) such that*

$$\begin{gathered}(S, T, G) \in \mathscr{B}(\boldsymbol{\Delta}), \quad S > 0, \quad T > 0, \\ \begin{bmatrix} D \\ I \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \rho^2 T & G \\ G^{\mathrm{T}} & -S \end{bmatrix} \begin{bmatrix} D \\ I \end{bmatrix} < 0.\end{gathered} \tag{3.2}$$

A lower bound on the well-posedness radius can be computed by solving the (generalized) semidefinite programming problem

$$\underline{\rho}^{\mathrm{wp}}(\boldsymbol{M}, \boldsymbol{\Delta}) = \sup \rho \quad \textit{subject to } (3.2). \tag{3.3}$$

Condition (3.2) *is also necessary in the unstructured case ($\boldsymbol{\Delta} = \mathbb{R}^{p\times q}$), in which case the well-posedness radius is $\rho^{\mathrm{wp}}(\boldsymbol{M}, \boldsymbol{\Delta}) = \underline{\rho}^{\mathrm{wp}}(\boldsymbol{M}, \boldsymbol{\Delta}) = \|D\|^{-1}$ if $D \neq 0$, and infinite otherwise.*

Proof. See Appendix B. □

3.2. Lower bound on invertibility radius

The matrix function $\mathbf{M}(\Delta) = \mathbf{A}(\Delta)^{-1}$ admits the LFR

$$\mathbf{M}(\Delta) = \mathbf{A}(\Delta)^{-1} = \tilde{M} + \tilde{L}\Delta(I - \tilde{D}\Delta)^{-1}\tilde{R}, \tag{3.4}$$

where

$$\left[\begin{array}{c|c} \tilde{M} & \tilde{L} \\ \hline \tilde{R} & \tilde{D} \end{array}\right] = \left[\begin{array}{c|c} A^{-1} & -A^{-1}L \\ \hline RA^{-1} & D - RA^{-1}L \end{array}\right]. \tag{3.5}$$

We seek a sufficient condition ensuring that $\mathbf{A}(\Delta)$ is well-posed and invertible for every $\Delta \in \boldsymbol{\Delta}_\rho$. Invertibility of $\mathbf{A}(\Delta)$ for every $\Delta \in \boldsymbol{\Delta}_\rho$ is guaranteed if the LFR of

$\mathbf{A}(\Delta)^{-1}$ given above is well-posed over $\boldsymbol{\Delta}_\rho$. According to Lemma 3.1, the LFR of $\mathbf{A}(\Delta)^{-1}$ is well-posed over $\boldsymbol{\Delta}_\rho$ if there exist S, G, T such that

$$\begin{aligned} &(S, T, G) \in \mathscr{B}(\boldsymbol{\Delta}), \quad S > 0, \quad T > 0, \\ &\begin{bmatrix} D - RA^{-1}L \\ I \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \rho^2 S & G \\ G^{\mathrm{T}} & -T \end{bmatrix} \begin{bmatrix} D - RA^{-1}L \\ I \end{bmatrix} < 0. \end{aligned} \tag{3.6}$$

A lower bound on the well-posedness radius of the LFR of $\mathbf{A}(\Delta)^{-1}$ is given by the solution of the (generalized) semidefinite programming problem

$$\underline{\rho}^{\mathrm{wp}}(\mathbf{A}^{-1}, \boldsymbol{\Delta}) = \sup \rho \quad \text{subject to (3.6)}. \tag{3.7}$$

Corollary 3.2. *A lower bound on the invertibility radius of* $\mathbf{A}(\Delta)$ *given in* (1.2) *is given by*

$$\underline{\rho}^{\mathrm{inv}}(\mathbf{A}, \boldsymbol{\Delta}) = \min\left(\underline{\rho}^{\mathrm{wp}}(\mathbf{A}, \boldsymbol{\Delta}), \underline{\rho}^{\mathrm{wp}}(\mathbf{A}^{-1}, \boldsymbol{\Delta})\right),$$

where $\underline{\rho}^{\mathrm{wp}}(\mathbf{A}, \boldsymbol{\Delta})$ *is defined in* (3.3) *and* $\underline{\rho}^{\mathrm{wp}}(\mathbf{A}^{-1}, \boldsymbol{\Delta})$ *in* (3.7). *In the unstructured case* ($\boldsymbol{\Delta} = \mathbb{R}^{p\times q}$), *the bound is exact, and given by*

$$\underline{\rho}^{\mathrm{inv}}(\mathbf{A}, \mathbb{R}^{p\times q}) = \min\left(\|D\|^{-1}, \|D - RA^{-1}L\|^{-1}\right),$$

with the convention that $\|M\|^{-1} = \infty$ *if the matrix M is zero.*

4. Structured maximal inversion error

In this section, we seek an upper bound on the structured absolute error defined in (1.3). We assume that $0 < \rho < \underline{\rho}^{\mathrm{inv}}(\boldsymbol{A}, \boldsymbol{\Delta})$.

4.1. Robustness lemma

We seek to guarantee a certain property for a given rational matrix-valued function, using the LFR and semidefinite programming.

Precisely, we consider again the linear-fractional function of a matrix variable $\Delta \in \mathbb{R}^{p\times q}$ given in (3.1). For a given real, symmetric matrix of appropriate size W, we seek a sufficient condition ensuring that the LFR above is well-posed, and in addition

$$\begin{bmatrix} \mathbf{M}(\Delta)^{\mathrm{T}} & I \end{bmatrix} W \begin{bmatrix} \mathbf{M}(\Delta) \\ I \end{bmatrix} < 0 \tag{4.1}$$

for every $\Delta \in \boldsymbol{\Delta}_\rho$, where $\rho > 0$ is given. (Here, $\mathbf{M}(\Delta)$ is a linear-fractional function as given in (3.1).) The motivation for studying this kind of condition is that the upper bound $\lambda(\mathbf{A}, \boldsymbol{\Delta}, \rho) < \lambda$ holds if and only if

$$\|\mathbf{A}(\Delta)^{-1} - X\| \leqslant \lambda \quad \text{for every } \Delta \in \boldsymbol{\Delta}_\rho,$$

which in turn can be expressed as (4.1) with $M(\Delta) = A(\Delta)^{-1} - A^{-1}$ and $W = \mathbf{diag}(I, -\lambda^2\rho^2 I)$.

We have the following result.

Lemma 4.1. *We have* $\det(I - D\Delta) \neq 0$ *and* (4.1) *for every* $\Delta \in \boldsymbol{\Delta}_\rho$, *if there exists a triple* (S, T, G) *such that* (3.2) *and*

$$\begin{bmatrix} M & L \\ I & 0 \end{bmatrix}^{\mathrm{T}} W \begin{bmatrix} M & L \\ I & 0 \end{bmatrix} + \begin{bmatrix} R & D \\ 0 & I \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \rho^2 S & G \\ G^{\mathrm{T}} & -T \end{bmatrix} \begin{bmatrix} R & D \\ 0 & I \end{bmatrix} < 0. \tag{4.2}$$

The condition is also necessary in the unstructured case ($\boldsymbol{\Delta} = \mathbb{R}^{p\times q}$), *in which case it can be expressed as* $\|D\| < \rho^{-1}$, *and there exists* $\tau \geqslant 0$ *such that*

$$\begin{bmatrix} M & L \\ I & 0 \end{bmatrix}^{\mathrm{T}} W \begin{bmatrix} M & L \\ I & 0 \end{bmatrix} + \begin{bmatrix} R & D \\ 0 & I \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \rho^2\tau I & 0 \\ 0 & -\tau I \end{bmatrix} \begin{bmatrix} R & D \\ 0 & I \end{bmatrix} < 0.$$

Proof . See Appendix C. □

The main implication of the above two lemmas is that a sufficient condition for both well-posedness and bound (4.1) can be checked using (generalized) semidefinite programming.

4.2. Upper bound on maximal inversion error

Applying Lemma 4.1, with $M(\Delta) = A^{-1}(\Delta) - A^{-1}$ and $W = \mathbf{diag}(I, -\lambda^2\rho^2 I)$, we obtain that the bound $\lambda(\mathbf{A}, \boldsymbol{\Delta}, \rho) < \lambda$ holds if there exists a triple $(S, T, G) \in \mathscr{B}(\boldsymbol{\Delta})$ such that $S > 0$, $T > 0$, and

$$\begin{bmatrix} \rho^2\lambda^2 I & 0 \\ 0 & 0 \end{bmatrix} > \begin{bmatrix} 0 \\ \tilde{L}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} 0 \\ \tilde{L}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} + \begin{bmatrix} \tilde{R} & \tilde{D} \\ 0 & I \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \rho^2 S & G \\ G^{\mathrm{T}} & -T \end{bmatrix} \begin{bmatrix} \tilde{R} & \tilde{D} \\ 0 & I \end{bmatrix}. \tag{4.3}$$

The above inequality implies that the LFR of $\mathbf{A}(\Delta)^{-1}$ is well-posed. Indeed, looking at the lower right corner of the above four-block inequality, we obtain

$$0 > \tilde{L}^{\mathrm{T}}\tilde{L} + \begin{bmatrix} \tilde{D} \\ I \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \rho^2 S & G \\ G^{\mathrm{T}} & -T \end{bmatrix} \begin{bmatrix} \tilde{D} \\ I \end{bmatrix}, \tag{4.4}$$

which implies (3.6).

Further, noting that

$$\begin{bmatrix} \tilde{R} & \tilde{D} \\ 0 & I \end{bmatrix} = \begin{bmatrix} RA^{-1} & D - RA^{-1}L \\ 0 & I \end{bmatrix} = \begin{bmatrix} R & D \\ 0 & I \end{bmatrix} N^{-1},$$

$$\text{where } N = \begin{bmatrix} A & L \\ 0 & I \end{bmatrix},$$

and multiplying inequality (4.3) on the left by N^{T} and on the right by N, we obtain a condition equivalent to (4.3):

$$\begin{bmatrix} \rho^2\lambda^2 A^{\mathrm{T}}A & \rho^2\lambda^2 A^{\mathrm{T}}L \\ \rho^2\lambda^2 L^{\mathrm{T}}A & L^{\mathrm{T}}(\rho^2\lambda^2 I - A^{-\mathrm{T}}A^{-1})L \end{bmatrix} > \begin{bmatrix} R & D \\ 0 & I \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \rho^2 S & G \\ G^{\mathrm{T}} & -T \end{bmatrix} \begin{bmatrix} R & D \\ 0 & I \end{bmatrix}. \tag{4.5}$$

(Note that, using Schur complements, it is possible to rewrite the above inequality in such a way that the computation of A^{-1} is not required.)

When $\Delta = \mathbb{R}^{p\times q}$, our condition is also necessary, and can be expressed as: there exists $\tau \geqslant 0$ such that

$$\begin{bmatrix} \rho^2\lambda^2 A^{\mathrm{T}}A & \rho^2\lambda^2 A^{\mathrm{T}}L \\ \rho^2\lambda^2 L^{\mathrm{T}}A & L^{\mathrm{T}}(\rho^2\lambda^2 I - A^{-\mathrm{T}}A^{-1})L \end{bmatrix} > \begin{bmatrix} R & D \\ 0 & I \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \rho^2\tau I & 0 \\ 0 & -\tau I \end{bmatrix} \begin{bmatrix} R & D \\ 0 & I \end{bmatrix}. \tag{4.6}$$

To find the maximal inversion error in the unstructured case, it suffices to do a line search over the parameter τ over the range $[\tau_{\mathrm{lb}}\ \infty]$, where

$$\tau_{\mathrm{lb}} = \|\tilde{L}(I - \rho^2\tilde{D}^{\mathrm{T}}\tilde{D})^{-1/2}\|^2.$$

For each value of τ in the interval $[\tau_{\mathrm{lb}}\ \infty]$, the corresponding minimal value of λ is given by the convex function

$$\lambda(\tau)^2 = \tau\lambda_{\max}\left(\tilde{R}^{\mathrm{T}}K(\tau)\tilde{R}\right), \quad \text{where } K(\tau) = I + \rho^2\tau\tilde{D}\left(\tau(I - \rho^2\tilde{D}^{\mathrm{T}}\tilde{D}) - \tilde{L}^{\mathrm{T}}\tilde{L}\right)\tilde{D}^{\mathrm{T}}. \tag{4.7}$$

We summarize the result as follows.

Theorem 4.2. *The maximal error bound is bounded above by*

$$\bar{\lambda}(\boldsymbol{A}, \rho) = \inf\lambda \quad \textit{subject to} \quad (4.5)\ \textit{and}\ (3.2).$$

When $\Delta = \mathbb{R}^{p\times q}$, the bound is exact, and can be expressed as the smallest value of the convex function $\lambda(\tau)$ given in (4.7)*, over the interval* $[\|\tilde{L}(I-\rho^2\tilde{D}^{\mathrm{T}}\tilde{D})^{-1/2}\|^2\ \infty]$.

5. Structured condition number

5.1. Upper bound

For every Δ such that the LFR (1.2) is well-posed, we have

$$\mathbf{A}(\Delta) - A^{-1} = \rho\tilde{L}\tilde{\Delta}(I - \rho\tilde{D}\tilde{\Delta})^{-1}\tilde{R},$$

where $\tilde{\Delta} = \Delta/\rho$, and $\tilde{L}$, $\tilde{R}$, $\tilde{D}$ are defined in (3.4). The above shows that

$$\kappa(\mathbf{A}, \Delta) = \limsup_{\rho \to 0} \frac{1}{\rho} \max_{\Delta \in \Delta_\rho} \|\mathbf{A}(\Delta)^{-1} - A^{-1}\| = \max_{\tilde{\Delta} \in \Delta_1} \|\tilde{L}\tilde{\Delta}\tilde{R}\|.$$

Redoing the derivation made in Section 4.2, with $\rho = 1$ and $\tilde{D} = 0$, we obtain that κ is an upper bound on the structured condition number if there exist (S, T, G) such that

$$(S, T, G) \in \mathscr{B}(\Delta), \quad S > 0, \quad T > 0,$$

$$\begin{bmatrix} \kappa^2 I & 0 \\ 0 & 0 \end{bmatrix} \geqslant \begin{bmatrix} 0 \\ \tilde{L}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} 0 \\ \tilde{L}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} + \begin{bmatrix} \tilde{R} & 0 \\ 0 & I \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} S & G \\ G^{\mathrm{T}} & -T \end{bmatrix} \begin{bmatrix} \tilde{R} & 0 \\ 0 & I \end{bmatrix}. \tag{5.1}$$

The above condition can be written

$$(S, T, G) \in \mathscr{B}(\Delta), \quad S > 0, \quad T > 0,$$

$$\begin{bmatrix} \kappa^2 A^{\mathrm{T}} A - R^{\mathrm{T}} S R & R^{\mathrm{T}} G^{\mathrm{T}} \\ GR & T - L^{\mathrm{T}} A^{-\mathrm{T}} A^{-1} L \end{bmatrix} \geqslant 0. \tag{5.2}$$

In the unstructured case ($\Delta = \mathbb{R}^{p \times q}$), our condition is also necessary, and is equivalent to the existence of a scalar τ such that

$$\kappa^2 A^{\mathrm{T}} A \geqslant \tau R^{\mathrm{T}} R, \quad \tau I \geqslant L^{\mathrm{T}} A^{-\mathrm{T}} A^{-1} L.$$

It is easy to show that the smallest value of κ^2 is obtained for $\tau^{\mathrm{opt}} = \|A^{-1}L\|^2$, and is equal to $\|RA^{-1}\|^2 \tau^{\mathrm{opt}}$.

Theorem 5.1. *The structured condition number is bounded above by* $\bar{\kappa}(\mathbf{A}, \Delta)$, *where*

$$\bar{\kappa}(\mathbf{A}, \Delta) = \inf \kappa^2 \text{ subject to } S > 0, \quad T > 0, \quad \text{and (5.2)}$$

When $\Delta = \mathbb{R}^{p \times q}$, *the bound is exact, and writes*

$$\bar{\kappa}(\mathbf{A}, \Delta) = \|A^{-1}L\| \cdot \|RA^{-1}\|. \tag{5.3}$$

In the additive case ($L = R = I$, $\Delta = \mathbb{R}^{p \times q}$), we recover the value of the absolute condition number, obtained from the classical result (1.1), namely, $\|A^{-1}\|^2$.

6. Approximate inverses

In this section, we again assume that $\rho < \underline{\rho}^{\mathrm{inv}}(\mathbf{A}, \Delta)$. Let $X \in \mathbb{R}^{n \times n}$. Applying Lemma 4.1, with $\mathbf{M}(\Delta) = A^{-1}(\Delta) - X$ and $W = \mathbf{diag}(I, -\lambda^2 \rho^2 I)$, we obtain that the bound

$$\|\mathbf{A}(\Delta)^{-1} - X\| < \lambda\rho$$

holds for every $\Delta \in \Delta_\rho$ if there exists a triple $(S, T, G) \in \mathscr{B}(\Delta)$ such that $S > 0$, $T > 0$, and

$$\begin{bmatrix} \rho^2 \lambda^2 I & 0 \\ 0 & 0 \end{bmatrix} > \begin{bmatrix} (A^{-1} - X)^{\mathrm{T}} \\ \tilde{L}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} (A^{-1} - X)^{\mathrm{T}} \\ \tilde{L}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$$

$$+\begin{bmatrix}\tilde{R} & \tilde{D}\\ 0 & I\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\rho^2 S & G\\ G^{\mathrm{T}} & -T\end{bmatrix}\begin{bmatrix}\tilde{R} & \tilde{D}\\ 0 & I\end{bmatrix}.$$

The above condition can be written, using Schur complements, as

$$\left[\begin{array}{cc|c} \multicolumn{2}{c|}{\tilde{\Theta}} & \begin{matrix}A^{-\mathrm{T}}-X^{\mathrm{T}}\\ \tilde{L}^{\mathrm{T}}\end{matrix}\\ \hline A^{-1}-X & \tilde{L} & I\end{array}\right] > 0,$$

where

$$\tilde{\Theta} = \begin{bmatrix}\rho^2\lambda^2 I & 0\\ 0 & 0\end{bmatrix} - \begin{bmatrix}\tilde{R} & \tilde{D}\\ 0 & I\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\rho^2 S & G\\ G^{\mathrm{T}} & -T\end{bmatrix}\begin{bmatrix}\tilde{R} & \tilde{D}\\ 0 & I\end{bmatrix}.$$

We now invoke the following lemma.

Lemma 6.1 (Elimination). *Let $A_{11}=A_{11}^{\mathrm{T}}$, A_{12}, $A_{22}=A_{22}^{\mathrm{T}}$, A_{23}, $A_{33}=A_{33}^{\mathrm{T}}$ be real matrices of appropriate size. There exists a matrix Y of appropriate size, such that*

$$\begin{bmatrix}A_{11} & A_{12} & Y\\ A_{12}^{\mathrm{T}} & A_{22} & A_{23}\\ Y & A_{23}^{\mathrm{T}} & A_{33}\end{bmatrix} > 0 \tag{6.1}$$

if and only if

$$\begin{bmatrix}A_{11} & A_{12}\\ A_{12}^{\mathrm{T}} & A_{22}\end{bmatrix} > 0, \text{ and } \begin{bmatrix}A_{22} & A_{23}\\ A_{23}^{\mathrm{T}} & A_{33}\end{bmatrix} > 0. \tag{6.2}$$

If condition (6.2) *holds, the set of Y's that satisfy* (6.1) *is parametrized by*

$$Y = A_{12}A_{22}^{-1}A_{23} + (A_{33}-A_{23}^{\mathrm{T}}A_{22}^{-1}A_{23})^{1/2}Z(A_{11}-A_{12}^{\mathrm{T}}A_{22}^{-1}A_{12})^{1/2},$$
$$\|Z\| < 1.$$

Proof . See Appendix D. □

Apply the elimination lemma to get an equivalent condition, namely (4.4) and $\tilde{\Theta} > 0$, that is,

$$0 > \tilde{L}^{\mathrm{T}}\tilde{L} + \begin{bmatrix}\tilde{D}\\ I\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\rho^2 S & G\\ G^{\mathrm{T}} & -T\end{bmatrix}\begin{bmatrix}\tilde{D}\\ I\end{bmatrix},$$

and

$$\begin{bmatrix}\rho^2\lambda^2 I & 0\\ 0 & 0\end{bmatrix} > \begin{bmatrix}\tilde{R} & \tilde{D}\\ 0 & I\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\rho^2 S & G\\ G^{\mathrm{T}} & -T\end{bmatrix}\begin{bmatrix}\tilde{R} & \tilde{D}\\ 0 & I\end{bmatrix}.$$

The above condition can be written

$$\rho^2\lambda^2\begin{bmatrix}A^{\mathrm{T}}\\ L^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}A^{\mathrm{T}}\\ L^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}} > \begin{bmatrix}R & D\\ 0 & I\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\rho^2 S & G\\ G^{\mathrm{T}} & -T\end{bmatrix}\begin{bmatrix}R & D\\ 0 & I\end{bmatrix} \text{ and (4.4).} \tag{6.3}$$

(The last inequality in the above implies that the LFR of $\mathbf{A}(\Delta)^{-1}$ is well-posed over $\boldsymbol{\Delta}_\rho$.) If S, G, T satisfy the above inequalities strictly, then a feasible X is given by

$$X = A^{-1} + A^{-1}L(\rho^2\tilde{D}^{\mathrm{T}}S\tilde{D} + \tilde{D}^{\mathrm{T}}G + G^{\mathrm{T}}\tilde{D} - S)^{-1} \times(\rho^2\tilde{D}^{\mathrm{T}}S + G^{\mathrm{T}})RA^{-1}. \tag{6.4}$$

In the unstructured case ($\Delta = \mathbb{R}^{p\times q}$), our condition is also necessary. The variables S, T are then proportional to τI_p and τI_q, respectively, where τ is a scalar. The approximate inverse, computed by specializing the expression above, turns out to be independent of the optimization variable τ:

$$X = A^{-1} + \rho^2 A^{-1}L(\rho^2\tilde{D}^{\mathrm{T}}\tilde{D} - I)^{-1}\tilde{D}^{\mathrm{T}}RA^{-1}, \quad \text{where } \tilde{D} = D - RA^{-1}L. \tag{6.5}$$

(We stress that the above analytic expression can be computed without any optimization.) In fact, it is easy to compute the expression for the optimal values of τ and λ in the unstructured case:

$$\tau^{\mathrm{opt}} = \|\tilde{L}(I - \rho^2\tilde{D}^T\tilde{D})^{-1/2}\|^2,$$
$$\lambda^{\mathrm{opt}} = \|\tilde{L}(I - \rho^2\tilde{D}^T\tilde{D})^{-1/2}\| \cdot \|(I - \rho^2\tilde{D}\tilde{D}^T)^{-1/2}\tilde{R}\|.$$

When $\rho = 0$, we recover—as expected—the expressions for the structured condition number (5.3).

Theorem 6.2. *A matrix X that minimizes the upper bound on the maximum error with respect to inversion is obtained by solving the SDP* (*in variables* λ^2, S, G, T *and X*):

$$\textit{minimize } \lambda \quad \textit{subject to } (6.3),\ (S, T, G) \in \mathscr{B}(\Delta), \quad S > 0, \quad T > 0.$$

In the unstructured case ($\Delta = \mathbb{R}^{p\times q}$), *the approximate inverse is given by the analytic expression* (6.5).

7. The additive case

Assume that the perturbation is additive, that is, the pertubed matrix A is of the form $\mathbf{A}(\Delta) = A + \Delta$, where $\Delta \in \mathbb{R}^{n\times n}$ satisfies $\|\Delta\| \leqslant \rho$, but is otherwise arbitrary. As said in Section 2, this kind of perturbation structure is a special case of the above, with $L = I_n$, $R = I_n$, $D = 0$. We will recover classical results in this case.

Introduce the SVD of A: $A = U\Sigma V^{\mathrm{T}}$, where $\Sigma = \mathbf{diag}(\sigma_1, \ldots, \sigma_n)$, and $\sigma_1 \geqslant \cdots \geqslant \sigma_n > 0$.

7.1. Invertibility radius

By application of Corollary 3.2, the invertibility radius is equal to $\rho = \|A^{-1}\|^{-1} = \sigma_n$, where σ_n is the smallest singular value of A. This is consistent with the fact that $A + \Delta$ is always well-posed, and invertible iff $\rho < \sigma_n$.

7.2. Maximal inversion error

We assume $\rho < \sigma_n$. In the additive perturbation case, the linear matrix inequality (4.6) writes

$$\begin{bmatrix} \lambda^2 A^{\mathrm{T}} A - \tau\rho^2 I & \lambda^2 A^{\mathrm{T}} \\ \lambda^2 A & (\lambda^2+\tau)I - A^{-\mathrm{T}}A^{-1} \end{bmatrix} \geqslant 0, \quad \tau \geqslant 0,$$

or, equivalently,

$$\tau \geqslant 0, \quad \begin{bmatrix} \lambda^2\sigma_i^2 - \tau\rho^2 & \lambda^2\sigma_i \\ \lambda^2\sigma_i & \lambda^2 + \tau - \sigma_i^{-2} \end{bmatrix} \geqslant 0, \quad i = 1, \ldots, n. \tag{7.1}$$

Condition (7.1) implies that $\lambda^2 + \tau - \sigma_i^{-2} > 0$ (otherwise, $\lambda\sigma_i = 0$). We obtain that condition (7.1) is equivalent to

$$\begin{aligned} &(\lambda^2\sigma_i^2 - \tau\rho^2)(\lambda^2 + \tau - \sigma_i^{-2}) \geqslant \lambda^4\sigma_i^2, \\ &\quad \tau \geqslant 0, \ \ \lambda^2 + \tau - \sigma_i^2 > 0, \ \ i = 1, \ldots, n. \end{aligned}$$

The first inequality is equivalent to

$$\lambda^2((\sigma_i^2 - \rho^2)\tau - 1) \geqslant \tau\rho^2(\tau - \sigma_i^2), \quad i = 1, \ldots, n.$$

Since $\tau(\sigma_i^2 - \rho^2) - 1 = 0$ for some i would imply $\lambda = \infty$ (which is ruled out by the invertibility of $A + \Delta$ whenever $\Delta \in \boldsymbol{\Delta}_\rho$), we finally obtain that the optimal value of λ satisfies

$$\lambda^2 = \max_{\tau \geqslant 0} \max_{1 \leqslant i \leqslant n} \quad \frac{1}{\sigma_i^2} \frac{\tau(\sigma_i^2\tau - 1)}{(\sigma_i^2 - \rho^2)\tau - 1}.$$

It is straightforward to show that the optimal value of τ is

$$\tau = \frac{1}{\sigma_n(\sigma_n - \rho)},$$

and the corresponding optimal value of λ is

$$\lambda(\mathbf{A}, \mathbb{R}^{p\times q}, \rho) = \frac{1}{\sigma_n(\sigma_n - \rho)}. \tag{7.2}$$

7.3. Link with the classical condition number

In the limit when $\rho \to 0$, we recover the classical absolute error bound:

$$\limsup_{\rho \to 0} \lambda(\mathbf{A}, \mathbb{R}^{p\times q}, \rho) = \frac{1}{\sigma_n^2} = \|A^{-1}\|^2.$$

We note that the *absolute* condition number, derived from the bound (1.1), is $\|A^{-1}\|^2$. Thus, we have recovered the classical condition number in this case.

7.4. Approximate inverse

In the additive case, we obtain a unique approximate inverse, given by

$$X(\rho) = (A^{\mathrm{T}}A - \rho^2 I)^{-1}A^{\mathrm{T}},$$

for which the maximal inversion error (as defined in (1.4)) is

$$\frac{1}{(\sigma_n^2 - \rho^2)}.$$

Comparing the above expression with that of the maximal inversion error (at A^{-1}), as given in (7.2), we conclude that choosing $X(\rho)$ instead of A^{-1} improves the maximal error bound by a relative amount of $\rho/(\sigma_n + \rho)$. This improvement increases when σ_n decreases (that is, as A becomes singular), and also when ρ increases.

The approximate inverse above comes up in the total least squares (TLS) problem. Precisely, the solution of the TLS problem

$$\text{minimize } \|[\Delta A \ \Delta b]\| \quad \text{subject to } (A + \Delta A)x = b + \Delta b,$$

where $b \in \mathbb{R}^n$ is given, is (except in degenerate cases):

$$x_{\mathrm{TLS}} = (A^{\mathrm{T}}A - \rho_{\mathrm{TLS}}^2)^{-1}A^{\mathrm{T}}b = X(\rho_{\mathrm{TLS}})b,$$

where $\rho_{\mathrm{TLS}} = \sigma_{\min}([A \ b])$.

The solution of the standard least squares (LS) problem, $\min_x \|Ax - b\|$, involves the classical inverse: $x_{\mathrm{LS}} = A^{-1}b$. The solution to the TLS problem is $x_{\mathrm{TLS}} = X(\rho_{\mathrm{TLS}})b$, where $X(\rho_{\mathrm{TLS}})$ is the approximate inverse of A. The TLS method amounts to first compute the smallest perturbation level necessary to make the linear system $Ax = b$ cosistent, ρ_{TLS}. Then, the TLS solution is computed via the approximate inverse (with level ρ_{TLS}), while the LS solution uses the standard inverse. This is coherent with the observation that "the TLS solution is more optimal than the LS one from a statiscal point of view" [1] when the matrix A is noisy. Indeed, the TLS solution works with an inverse matrix that best approximates the possible values of $(A + \Delta A)^{-1}$, over the smallest perturbation range making the system consistent.

8. Extensions

Our approach can be generalized to a number of related problems. Possible extensions are as follows.

8.1. Componentwise error bounds

In some applications, one may be interested in componentwise error bounds, instead of global (norm) bounds on the inversion error. We define the inversion error matrix $\Lambda(\mathbf{A}, \boldsymbol{\Delta}, \rho) = (\lambda_{ij}(\mathbf{A}, \boldsymbol{\Delta}, \rho))_{1 \leqslant i,j \leqslant n}$, where, for every i, j $(1 \leqslant i, j \leqslant n)$,

$$\lambda_{ij}(\mathbf{A}, \mathit{\Delta}, \rho) = \frac{1}{\rho} \max \left\{ \left| (\mathbf{A}(\mathit{\Delta})^{-1})_{ij} - (A^{-1})_{ij} \right| : \mathit{\Delta} \in \mathit{\Delta}_\rho \right\}.$$

(In the above definition, the notation $(M^{-1})_{ij}$ stands for the (i, j) element of M^{-1}.) We may define similarly a componentwise condition matrix. Finally, we may define an approximate inverse, as a matrix X which minimizes, e.g., the l_1 norm on the matrix whose i, j element is

$$\frac{1}{\rho} \max \left\{ \left| (\mathbf{A}(\mathit{\Delta})^{-1})_{ij} - X_{ij} \right| : \mathit{\Delta} \in \mathit{\Delta}_\rho \right\}.$$

(Note that the above matrix can be used as an approximation in interval arithmetic computations.)

It is straightforward to extend the previous approach to the above problem, once it is noted that an LFR of the matrix function $(\mathbf{A}(\mathit{\Delta})^{-1})_{ij}$ is given by

$$(\mathbf{A}(\mathit{\Delta})^{-1})_{ij} = e_i^{\mathrm{T}} A^{-1} e_j + e_i^{\mathrm{T}} \tilde{L} \mathit{\Delta} (I - \tilde{D}\mathit{\Delta})^{-1} \tilde{R} e_j,$$

where $\tilde{L}, \tilde{R}, \tilde{D}$ are given in (3.5).

8.2. Adding constraints on inverse and perturbation

Another possible extension is to add structure constraints on the approximate inverse if it is a priori known. Indeed, it may be judicious to impose additional (linear) constraints on the matrix X in the semidefinite program (6.4). For example, if the nominal matrix A is symmetric, we might want the approximate inverse to be symmetric as well. (It turns out that if the perturbed matrix $\mathbf{A}(\mathit{\Delta})$ is always symmetric, then so is the approximate inverse defined in (6.4). Therefore adding such linear constraints only makes sense when the perturbation changes the structure of the nominal matrix A.)

Another extension is to adapt the same methods to other kinds of bounds on the perturbation matrix $\mathit{\Delta}$, such as $\mathit{\Delta} + \mathit{\Delta}^{\mathrm{T}} \geqslant 0$, instead of the norm bound $\|\mathit{\Delta}\| \leqslant \rho$.

9. Example: the Vandermonde case

The following numerical experiment was performed using the SDP code **SP** [26] and a preprocessor called `lmitool` [8].

We consider the problem of inverting a matrix $\mathbf{A}(a)$ with Vandermonde structure, as defined in (2.6). The nominal value of the vector a is chosen to be

$$a = [1\ 1.2\ 2.5\ 3.1].$$

The perturbed matrix $\mathbf{A}(a + \delta)$ can be expressed with the LFR (1.2), where $A = \mathbf{A}(a)$, and L, R, D and $\mathit{\Delta}$ given in Section 2.4. This LFR is always well-posed: $\det(I - D\mathit{\Delta}) \neq 0$ for every diagonal $\mathit{\Delta}$, since D is stricly upper triangular. For Vandermonde matrices, the invertibility radius is easy to compute, and is given by (1/2)

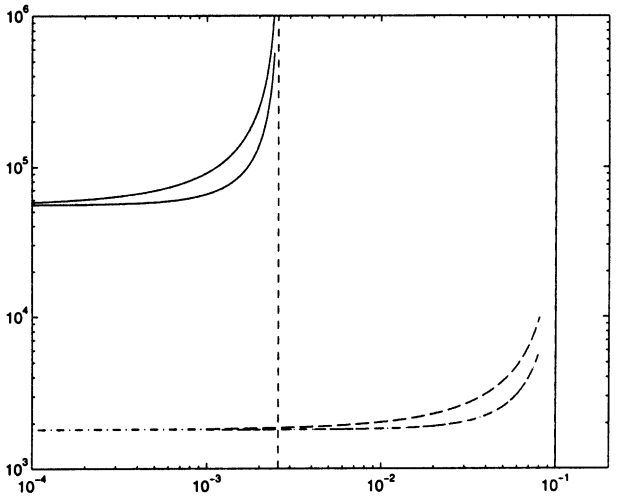

Fig. 1. Comparison of unstructured maximal inversion error (top curve) and optimized unstructured maximal inversion error (curve below top), with their structured counterparts (next-to-bottom and bottom curves, respectively), as functions of the perturbation level ρ. The right and left vertical lines correspond to the unstructured and structured invertibility radiuses, respectively.

$\min_{i\neq j} |a_i - a_j|$. We have computed a lower bound on this exact value by finding the largest ρ such that condition (3.6) is feasible. Our estimate is 0.0995, whereas the exact value is 0.1. We note that if we neglect the structure information, we obtain the (exact value of) the unstructured invertibility radius, which is considerably smaller than the structured counterpart: 0.0026. (Our results are hard to compare with a purely classical analysis, where the perturbation size is measured in terms of additive errors in the matrix A; our analysis is based on errors described by the perturbation matrix Δ only.)

In Fig.1, we show the maximal inversion error as a function of the perturbation level ρ. In order to illustrate the importance of structure in the inversion problem, we show two pairs of curves. The top curves unstructured maximal inversion error and to the optimized unstructured maximal inversion error, respectively. The bottom curves refer to the structured couterpart. The plot also shows two vertical lines. The right line corresponds to the unstructured invertibility radius, the left one to its structured counterpart.

The plot shows clearly that neglecting structure leads to the conclusion that the Vandermonde matrix at hand is very ill-conditioned (note the logarithmic scale). The structured analysis is more consistent with the actual conditioning of the system, which much better than its unstructured counterpart. Another interesting point to observe is the improvement brought by using approximate inverses instead of the classical inverse; the improvement grows with ρ. Thus, using an approximate inverse

instead of the classical one makes more and more sense as the perturbation level grows.

10. Concluding remarks

In this paper we have proposed an approach to rigorously measure, and reduce the effect, of possibly large, structured perturbations in the computation of an inverse matrix. In a future paper we will investigate a similar approach for the solution of a general structured linear system.

Acknowledgments

Useful comments of the reviewers are gratefully acknowledged.

Appendix A. Invariance with respect to LFR model

In this section, we show that the sufficient conditions obtained in this paper are, in some sense, independent of the LFR model used to describe the perturbation structure.

First, note that our results are based on an LFR of a function taking values in the set of symmetric matrices $\mathbf{F}(\Delta)$ having an LFR such as

$$\mathbf{F}(\Delta) = F + L\Delta(I - D\Delta)^{-1}R + (L\Delta(I - D\Delta)^{-1}R)^{\mathrm{T}}. \tag{A.1}$$

(For example, the condition $\|\mathbf{A}(\Delta)^{-1} - X\| \leqslant \lambda$ can be written $\mathbf{F}(\Delta) \geqslant 0$ for appropriate $\mathbf{F}$.)

Now, consider a function taking values in the set of symmetric matrices having an LFR such as (A.1). This function can be written in a more symmetric form

$$\mathbf{F}(\Delta) = F + \tilde{L}\tilde{\Delta}(I - D\tilde{\Delta})^{-1}\tilde{L}^{\mathrm{T}}, \tag{A.2}$$

where

$$\tilde{L} = \begin{bmatrix} L & R^{\mathrm{T}} \end{bmatrix}, \quad \tilde{D} = \begin{bmatrix} 0 & D^{\mathrm{T}} \\ D & 0 \end{bmatrix}, \quad \tilde{\Delta} = \begin{bmatrix} 0 & \Delta \\ \Delta^{\mathrm{T}} & 0 \end{bmatrix}.$$

It is easy to check that, if an invertible matrix Z satisfies the relation $Z\tilde{\Delta}Z^{\mathrm{T}} = \tilde{\Delta}$ for every $\Delta \in \boldsymbol{\Delta}$, then

$$\mathbf{F}(\Delta) = F + (\tilde{L}Z)\tilde{\Delta}(I - (Z^{\mathrm{T}}\tilde{D}Z)\tilde{\Delta})^{-1}(\tilde{L}Z)^{\mathrm{T}}.$$

In other words, the "scaled" triple $(F, (\tilde{L}Z), (Z^{\mathrm{T}}\tilde{D}Z))$ can be used to represent $\mathbf{F}$ instead of $F, \tilde{L}, \tilde{D}$ in (A.2).

A valid scaling matrix Z can be constructed as follows. Let $(S, T, G) \in \mathscr{B}(\boldsymbol{\Delta})$, and define

$$Z = \begin{bmatrix} T^{-1/2} & 0 \\ 0 & S^{1/2} \end{bmatrix} \begin{bmatrix} I & G \\ 0 & I \end{bmatrix}.$$

It turns out that such an Z satisfies the relation $Z\tilde{\Delta}Z^{\mathrm{T}} = \tilde{\Delta}$ for every $\Delta \in \boldsymbol{\Delta}$. It turns out that the conditions we obtained (e.g. , in Lemma 4.1) amount to a search over the scaling matrix Z. In this sense, our conditions are independent of the LFR used to represent the perturbation structure.

Appendix B. Proof of Lemma 3.1

We first recall a well-known lemma. A proof can be found e.g. in [4, p. 24]. This lemma is widely used, e.g. in control theory, and in connection with trust region methods in optimization [25].

Lemma B.1 ($\mathscr{S}$-procedure). *Let F_0, F_1 be quadratic functions of the variable $\xi \in \mathbb{R}^m$:*

$$F_i(\xi) \stackrel{\Delta}{=} \xi^{\mathrm{T}} T_i \xi + 2u_i^{\mathrm{T}} \xi + v_i, \quad i = 0, 1,$$

where $T_i = T_i^{\mathrm{T}}$. The following condition on F_0, F_1:

$$F_0(\xi) \geqslant 0 \quad \textit{for all } \xi \textit{ such that } F_1(\xi) \geqslant 0$$

holds if

$$\textit{there exist } \tau \geqslant 0 \quad \textit{such that} \quad \begin{bmatrix} T_0 & u_0 \\ u_0^{\mathrm{T}} & v_0 \end{bmatrix} \geqslant \tau \begin{bmatrix} T_1 & u_1 \\ u_1^{\mathrm{T}} & v_1 \end{bmatrix}.$$

The converse holds, provided that there is some ξ_0 such that $F_1(\xi_0) > 0$.

We will also use the fact that if a triple (S, T, G) belongs to $\mathscr{B}(\boldsymbol{\Delta})$, with $S > 0$ and $T > 0$, then for every ξ, ζ such that $\xi = \Delta\zeta$ for some $\Delta \in \boldsymbol{\Delta}$, $\|\Delta\| \leqslant \rho$, we have $\rho^2 \zeta^{\mathrm{T}} T \zeta \geqslant \xi^{\mathrm{T}} S \xi$ and $\zeta^{\mathrm{T}} G \xi = 0$. The latter can be written compactly as

$$\begin{bmatrix} \zeta \\ \xi \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \rho^2 S & G \\ G^{\mathrm{T}} & -T \end{bmatrix} \begin{bmatrix} \zeta \\ \xi \end{bmatrix} \geqslant 0. \tag{B.1}$$

The above outer approximation is exact in the unstructured case ($\boldsymbol{\Delta} = \mathbb{R}^{\xi \times q}$).

We proceed to prove Lemma 3.1. Note that $\det(I - \Delta D) \neq 0$ for every $\Delta \in \boldsymbol{\Delta}$, $\|\Delta\| \leqslant \rho$, if and only if $\xi = 0$ whenever $\zeta = D\xi$, $\xi = \Delta\zeta$, $\Delta \in \boldsymbol{\Delta}$, $\|\Delta\| \leqslant \rho$. This property is equivalent to

$$\|\xi\|^2 \leqslant 0 \quad \text{whenever } \xi = \Delta\zeta,\ \Delta \in \boldsymbol{\Delta},\ \|\Delta\| \leqslant \rho, \text{ and } \zeta = D\xi.$$

The above is true if $\|\xi\|^2 \leqslant 0$ whenever (B.1). We now apply the $\mathscr{S}$-procedure above, with

$$F_0(\xi) = -\|\xi\|^2, \quad F_1(\xi) = \begin{bmatrix} D\xi \\ \xi \end{bmatrix} \begin{bmatrix} \rho^2 T & G \\ G^{\mathrm{T}} & -S \end{bmatrix} \begin{bmatrix} D\xi \\ \xi \end{bmatrix}.$$

We obtain a sufficient condition for well-posedness: there exists a scalar $\tau \geqslant 0$ such that

$$\text{for every } \xi, \ F_0(\xi) \geqslant \tau F_1(\xi).$$

We absorb the scalar in the matrices S, G, T and obtain that (3.2) is a sufficient condition for well-posedness.

To check that condition (3.2) is also necessary in the unstructured case $\boldsymbol{\Delta} = \mathbb{R}^{p\times q}$, it suffices to recall the structure of the set $\mathscr{B}(\boldsymbol{\Delta})$ in this case (given in (2.2)). Condition (3.2) is equivalent to $\rho^2 D^{\mathrm{T}} D < I$, which is indeed equivalent to the invertibility of $I - D\Delta$ for every Δ, $\|\Delta\| \leqslant \rho$.

Appendix C. Proof of Lemma 4.1

To prove Lemma 4.1, assume the LFR of $\mathbf{M}$ is well-posed over $\boldsymbol{\Delta}_\rho$. Constraint (4.1) holds for every $\Delta \in \boldsymbol{\Delta}_\rho$ if and only if

$$\begin{bmatrix} u \\ \xi \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} M & L \\ I & 0 \end{bmatrix}^{\mathrm{T}} W \begin{bmatrix} M & L \\ I & 0 \end{bmatrix} \begin{bmatrix} u \\ \xi \end{bmatrix} \leqslant 0$$

for every u, ξ such that $\xi = \Delta(Ru + D\xi)$ for some $\Delta \in \boldsymbol{\Delta}$, $\|\Delta\| \leqslant \rho$. As seen in Appendix B, the latter condition implies that for every every triple $(S, T, G) \in \mathscr{B}(\boldsymbol{\Delta})$, we have

$$\begin{bmatrix} u \\ \xi \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} R & D \\ 0 & I \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \rho^2 S & G \\ G^{\mathrm{T}} & -T \end{bmatrix} \begin{bmatrix} R & D \\ 0 & I \end{bmatrix} \begin{bmatrix} u \\ \xi \end{bmatrix} \geqslant 0.$$

The conclusions of the lemma follow by a straightforward application of the $\mathscr{S}$-procedure lemma. In the unstructured case, we recover a necessary and sufficient condition, as in Appendix B.

Appendix D. Proof of Lemma 6.1

The condition is obviously necessary. Now if (6.2) holds, then A_{22} is invertible. Using Schur complements, we rewrite (6.1) as

$$\begin{bmatrix} A_{11} & Y \\ Y & A_{33} \end{bmatrix} - \begin{bmatrix} A_{12} \\ A_{23}^{\mathrm{T}} \end{bmatrix} A_{22}^{-1} \begin{bmatrix} A_{12}^{\mathrm{T}} & A_{23} \end{bmatrix} > 0,$$

or, equivalently,

$$\begin{bmatrix} \tilde{A}_{11} & Y - Y_0 \\ (Y - Y_0)^{\mathrm{T}} & \tilde{A}_{33} \end{bmatrix} > 0, \tag{D.1}$$

where $Y_0 = A_{12}A_{22}^{-1}A_{23}$, $\tilde{A}_{11} = A_{11} - A_{12}A_{22}^{-1}A_{12}^{\mathrm{T}}$, $\tilde{A}_{33} = A_{33} - A_{23}^{\mathrm{T}}A_{22}^{-1}A_{23}$. (We note that $\tilde{A}_{11}$ and $\tilde{A}_{33}$ are both invertible when (6.2) holds.) To conclude, it suffices to introduce the variable $Z = \tilde{A}_{11}^{-1/2}(Y - Y_0)\tilde{A}_{33}^{-1/2}$, and express (D.1) equivalently as $\|Z\| < 1$.

References

[1] T.J. Abatzoglou, J.M. Mendel, G.H. Harada, The constrained total least squares technique and its applications to harmonic resolution, IEEE Trans. Signal Processing 39 (1991) 1070–1087.
[2] S.G. Bartels, D.J. Higham, The structured sensitivity of Vandermonde-like systems, Numer. Math. 62 (1992) 17–33.
[3] S. Boyd, L. El Ghaoui, Method of centers for minimizing generalized eigenvalues in: Numerical Linear Algebra Methods in Control, Signals and Systems (special issue), Linear Algebra Appl. 188 (1993).
[4] S. Boyd, L. El Ghaoui, E. Feron, V. Balakrishnan, Linear Matrix Inequalities in System and Control Theory, Studies in Applied Mathematics, vol. 15, SIAM, Philadelphia, PA, June 1994.
[5] P. Comon, Structured matrices and inverses, Linear Algebra for Signal Processing, IMA, vol. 69 (1995).
[6] J. Demmel, The componentwise distance to the nearest singular matrix, SIAM J. Matrix Anal. Appl. vol. 13 (1992) 10–19.
[7] L. El Ghaoui, H. Lebret, Robust solutions to least-squares problems with uncertain data, SIAM J. Matrix Anal. Appl. 18 (1997) 1035–1064.
[8] L. El Ghaoui, R. Nikoukhah, F. Delebecque, LMITOOL: A front-end for LMI optimization, user's guiide, February 1995. Available via anonymous ftp to ftp.ensta.fr, under /pub/elghaoui/lmitool.
[9] M.K.H. Fan, A.L. Tits, J.C. Doyle, Robustness in the presence of mixed parametric uncertainty and unmodeled dynamics, IEEE Trans. Aut. Control 36 (1991) 25–38.
[10] I. Gohberg, T. Kailath, V. Olshevsky, Fast Gaussian elimination with partial pivoting for matrices with displacement structure, Math. Comput. 64 (1995) 1557–1576.
[11] I. Gohberg, I. Koltracht, Structured condition numbers for linear matrix structures, Linear Algebra for Signal Processing, IMA, vol. 69 (1995).
[12] W. Hagher, Condition estimates, SIAM J. Sci. Statist. Comput. 5 (1984) 311–316.
[13] D.J. Higham, N.J. Higham, Backward error and condition of structured linear systems, SIAM J. Matrix Anal. Appl. 13 (1992) 162–175.
[14] N.J. Higham, Accuracy and Stability of Numerical Algorithms, SIAM, Philadelphia, PA, 1996.
[15] R.E. Moore, Interval Analysis, Prentice-Hall, Englewood Cliffs, NJ, 1966.
[16] R.E. Moore, On computing the range of values of a rational function of *n* variables over a bounded region, Computing 16 (1976) 1–15.
[17] A. Nemirovsky, Several NP-hard problems arising in robust stability analysis, MCSS 6 (1993) 99–105.
[18] Y. Nesterov, A. Nemirovskii, An interior-point method for generalized linear-fractional problems, Math. Programming, Ser. B (1993).
[19] Y. Nesterov, A. Nemirovsky, Interior Point Polynomial Methods in Convex Programming: Theory and Applications, SIAM, Philadelphia, PA, 1994.
[20] S. Poljak, J. Rohn, Checking robust nonsingularity is NP-hard, Math. Control Signals Systems 6 (1993) 1–9.
[21] J. Rohn, Systems of linear interval equations, Linear Algebra Appl. 126 (1989) 39–78.
[22] J. Rohn, Nonsingularity under data rounding, Linear Algebra Appl. 139 (1990) 171–174.
[23] J. Rohn, Overestimations in bounding solutions of perturbed linear equations, Linear Algebra Appl. 262 (1997) 55–66.

[24] S.M. Rump, Bounds on the componentwise distance to the nearest singular matrix, SIAM J. Matrix Anal. Appl. 18 (1997) 83–103.
[25] R.J. Stern, H. Wolkowicz, Indefinite trust region subproblems and nonsymmetric eigenvalue perturbations, SIAM J. Optim. 5 (1995) 286–313.
[26] L. Vandenberghe, S. Boyd, SP, Software for semidefinite programming, user's guide, December 1994. Available via anonymous ftp to isl.stanford.edu under /pub/boyd/semidef_prog.
[27] L. Vandenberghe, S. Boyd, Semidefinite programming, SIAM Rev. 38 (1996) 49–95.
[28] K. Zhou, J. Doyle, K. Glover, Robust and Optimal Control, Prentice-Hall, Englewood Cliffs, NJ, 1995.