SIAM J. MATRIX ANAL. APPL.
Vol. 23, No. 1, pp. 138–166

THE DEGENERATE BOUNDED ERRORS-IN-VARIABLES MODEL*

S. CHANDRASEKARAN†, M. GU‡, A. H. SAYED§, AND K. E. SCHUBERT¶

Abstract. We consider the following problem: $\min_{x\in\mathcal{R}^n} \min_{\|E\|\le\eta} \|(A+E)x-b\|$, where A is an $m\times n$ real matrix and b is an n-dimensional real column vector when it has multiple global minima. This problem is an errors-in-variables problem, which has an important relation to total least squares with bounded uncertainty. A computable condition for checking if the problem is degenerate as well as an efficient algorithm to find the global solution with minimum Euclidean norm are presented.

Key words. least squares problems, total least squares, errors-in-variables, parameter estimation

AMS subject classifications. 65F05, 65F20, 65F25, 65G05, 93B40, 93E24

PII. S0895479800357766

1. Introduction.

In this paper we consider the following problem:

$$\min_{x\in\mathcal{R}^n} \min_{\|E\|\le\eta} \|(A+E)x-b\|, \tag{1.1}$$

where A is an $m\times n$ real matrix and b is a real n-vector. This problem is a special case of the errors-in-variables problem, which we have given the formal name of the degenerate bounded errors-in-variables problem. For ease of reference we usually call the problem the degenerate min-min problem, since degenerate bounded errors-in-variables problem is a bit long. This problem can be viewed as a total least squares (TLS) problem [3, 4] with bounds on the uncertainty in the coefficient matrix, which we will explain in more detail in section 3. In this paper we make frequent use of the terms degenerate and nondegenerate. Simply put, a degenerate problem is one where multiple solutions exist. The nondegenerate case of this problem occurs when η is small and b is in some sense far from the range of A. That η should be small is intuitive, since for $\eta = 0$ we are left with the least squares problem, which is nondegenerate (unique solution) when A has full column rank. Conversely, when η is larger than the smallest singular value of A, we would anticipate degeneracy (multiple solutions) as the perturbed matrix $A+E$ is not guaranteed to be full column rank. The intuition behind b needing to be far from the range of A for nondegeneracy comes from the fact that if b were close enough that multiple perturbations E existed such that b was in the range of $A+E$, then multiple solutions (degeneracy) would exist. In [2] we considered the nondegenerate case of this problem and showed how to

*Received by the editors April 7, 2000; accepted for publication (in revised form) by S. Van Huffel August 18, 2000; published electronically June 8, 2001. The first and fourth authors were partially supported by NSF grant CCR-9734290. The third author was supported in part by the NSF under award CCR-9732376.

http://www.siam.org/journals/simax/23-1/35776.html

†Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 (shiv@ece.ucsb.edu).

‡Department of Mathematics, University of California, Los Angeles, CA 90095 (mgu@math.ucla.edu).

§Department of Electrical Engineering, University of California, Los Angeles, CA 90095 (sayed@ee.ucla.edu).

¶Mathematics Department, University of Redlands, Redlands, CA 92373-0999 (schubert@jasper.uor.edu).

compute its unique solution in $O(mn^2)$ flops. In this paper we consider the problem when it is degenerate; that is, when it has multiple solutions. In particular, we present an $O(mn^2)$ algorithm to find the solution with the minimum Euclidean norm. The degenerate case is actually the generic case for this problem, and hence is more important than the nondegenerate case. This can be seen from the simple discussion above, since the nondegenerate case holds only for certain combinations of b and A when η is smaller than the smallest singular value of A. This is very restrictive, and hence the claim.

We begin this paper with a motivational problem, which shows the advantage of considering this criterion. We proceed by outlining the proof and presenting the algorithm to solve the problem. We then proceed with the full proof of the problem. We conclude with a tabulation of the results and an extension to the problem of a column partitioned matrix with uncertainty in only one partition.

2. Motivation. Many different methods exist for solving the basic estimation problem of finding some vector of unknowns x, from a vector of observations b, by using a matrix of relations A. Probably the two best known methods are least squares and TLS. We now want to get a feel for how these problems operate on a simple example and to see if there is any room for improvement. Consider, for example, a simple one dimensional "skyline" image that has been blurred. A "skyline" image is a one dimensional image that looks like a city skyline when graphed, and thus is the most basic image processing example. "Skyline" images involve sharp corners, and it is of key importance to accurately locate these corner transitions. Blurring occurs often in images; for example, atmospheric conditions, dust, or imperfections in the optics can cause a blurred image. Blurring is usually modeled as a Gaussian function or Gaussian blur, which incidentally is a great smoothing filter. The Gaussian blur causes greater distortion on the corners, which is exactly where we do not want it to happen. The Gaussian blur with standard deviation, σ, can be modeled as a matrix, A, with the component in position, (i,j), given by

$$A_{i,j} = e^{-(i-j)^2\sigma}.$$

If we go on the presumption that we do not know the exact blur that was applied (σ unknown) we cannot expect to get the exact system back. We realize that we will not be able to perfectly extract the original system, but we want to see if we can get a little more information than we have now. We "know" the blur is small compared to the information so we are confident that we should be able to get something. The least squares solution fails completely, yielding a result that is about three orders of magnitude off; see Figure 1. We notice that the TLS solution is better than the least squares solution, but still not acceptable. The degenerate min-min problem yields great results. From this simple example we can see that there is room for improvement.

3. Geometric understanding. Probably the easiest way to understand the problem at hand is to look at it geometrically. For ease of drawing we will consider A and b to be vectors of length 2. Note that while this is useful for getting a basic understanding some of the key features of the problem do not appear in this case. For instance, when A has multiple columns the problem can be degenerate for small values of η. In such a case the degenerate min-min problem has several advantages over other formulations, such as TLS. One such advantage is the perturbation on A is much smaller in the degenerate min-min problem than in the TLS problem.

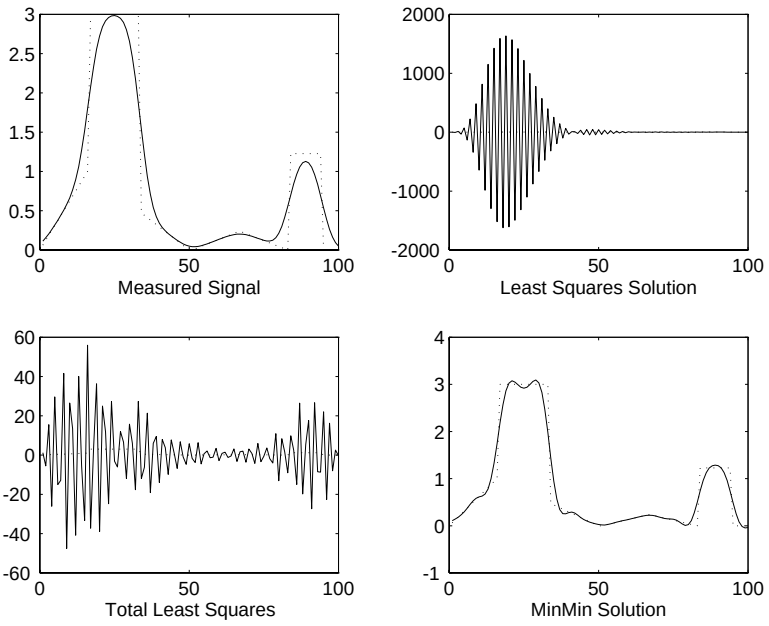

FIG. 1. *Skyline problem.*

For comparison we start with the classic problem of least squares (see Figure 2). The solution to the least squares problem is found by projecting b into A. This is a common geometric view of the problem, but forms a basis for understanding the other problems.

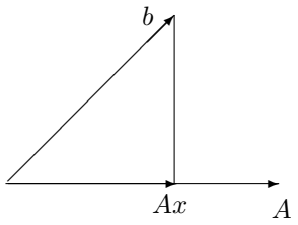

FIG. 2. *Least squares.*

In TLS, we allow A to be perturbed by a matrix E and b to be perturbed by a vector f (see Figure 3). The net effect is that both A and b are projected into a plane between the two such that the norm of $[\,E \quad f\,]$ is minimized. The TLS problem can thus be formulated as $\min \|[\,E \quad f\,]\|$ such that $(b+f) \in \mathcal{R}(A+E)$. Note that because of this A can be moved arbitrarily far.

In the general min-min problem (degenerate or not), we project A and b into a plane between the two as we did in the TLS problem, but we put a bound on how far A can be perturbed (see Figure 4). Note that the cone around A shows us the boundary of possible perturbations to A. We are in essence solving the problem $\min \|[\,E \quad f\,]\|$ such that $(b+f) \in \mathcal{R}(A+E)$ and $\|E\| \leq \eta$. The problem at hand can thus be thought of as a TLS problem with bounds on the errors in A.

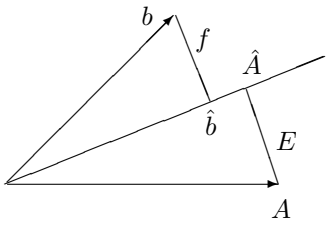

FIG. 3. *Total least squares (TLS).*

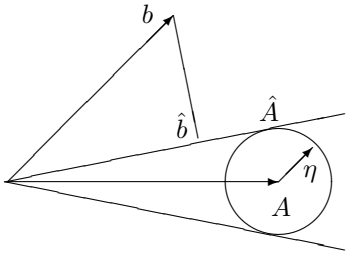

FIG. 4. *Min-min problem.*

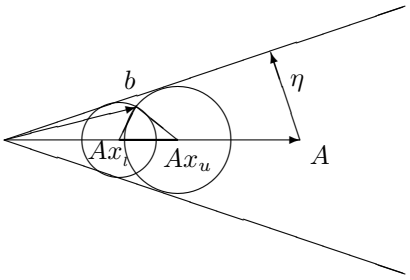

FIG. 5. *Degenerate min-min problem.*

To get a better understanding of the degenerate problem, we will consider one of the ways the problem can become degenerate. The easiest to visualize, and the only one that can be drawn in two dimensions, is the case when b lies in the cone of possible perturbations of A (see Figure 5). In this case we see that any $\hat{x}$ such that $x_l \leq \hat{x} \leq x_u$ is a solution to the problem. The perturbations $E(\hat{x})$ change, but each $\hat{x}$ in the range still solves the problem. We are now left with a problem, namely, which $\hat{x}$ do we choose. The most conservative choice is to pick the smallest one, which is what we do. This choice has a lot to recommend it, but a full discussion is outside the bounds of the paper at hand. In section 6, we take advantage of this basic insight (picking the smallest solution) to reformulate the problem into a unique problem.

4. Proof outline. The proof is long and technically involved, so we provide this overview. The cost function presented is useful for seeing how this problem handles the uncertainty in the matrix A, but it is not immediately useful in solving the problem. For instance, checking if a problem is degenerate in the original form of the problem

is tedious. We thus desire to rewrite the problem into a simpler form, and then find a computable condition for degeneracy. We start the proof with the cost function of the problem we want to solve,

$$\min_{x\in\mathcal{R}^n} \min_{\|E\|\leq\eta} \|(A+E)x-b\|,$$

and the degeneracy condition which was found in [2],

$$\eta\|x\| \geq \|Ax-b\|.$$

Since the nondegenerate case is already solved, we proceed by assuming the degeneracy condition holds. The first step is to minimize the cost function over $\|E\| \leq \eta$, and find that the optimal cost is zero. Since the problem is degenerate and the cost function is zero, we choose the solution with the smallest norm to obtain the problem

$$\min_{\|Ax-b\|\leq\eta\|x\|} \|x\|.$$

The condition $\eta\|x\| \geq \|Ax-b\|$ is not practical for checking for degeneracy in a problem, as mentioned above, since it requires the checking multiple values of x to hopefully find one that holds and thus showing the problem is degenerate. The second step is thus to find a computable condition for degeneracy. We proceed by squaring the condition for degeneracy and using the singular value decomposition (SVD) of A to find the two cases in which the problem is degenerate. The first case is when η is larger than the smallest singular value of A. The first case is always degenerate. The second case is when η is not larger than the smallest singular value of A. The second case is degenerate only when

$$b^T(I - A(A^TA - \eta^2 I)^{-1}A^T)b \leq 0.$$

While we now know when the problem is degenerate, we still need to show how to get the solution. We would like to be able to use Lagrange multiplier techniques to find the solution. We thus need to reduce the inequality $\eta\|x\| \geq \|Ax-b\|$ to an equality if possible. The third step of the proof is a proof that the solution, $\hat{x}$, is actually on the boundary of the inequality, and thus $\eta\|\hat{x}\| = \|A\hat{x}-b\|$.

We then proceed in the fourth step to use Lagrange multiplier techniques to parameterize the solution, $\hat{x} = x(\alpha)$, in terms of a single variable, α, thus reducing the problem to finding the zeros a secular equation. A secular equation is a rational expression of one variable, which we construct so that all the critical points of the original problem occur at zeros of the secular equation. The secular equation reduces our n-dimensional search for the solution, $\hat{x}$, to a one dimensional search. We denote the solution to the original problem as $x(\alpha^o)$, and note that it will occur at one of the $2n$ zeros of the secular equation. The zero of the secular equation which corresponds to $x(\alpha^o)$ is denoted α^o.

The remainder of the proof is concerned with showing which zero is α^o. Toward this end we start the second half of the proof with an assertion of the answer. The unique zero of the secular equation in the interval $[\max(-\sigma_n^2, -\eta^2), \eta\sigma_1]$ is α^o, where σ_1 is the largest singular value of A and σ_n is the smallest. We prove this by a process of elimination.

To begin with we us Lagrange techniques (first and second order conditions on the Lagrangian) to narrow down the search area. By employing these techniques, we

find that α^o must lie in the interval $[\max(-\sigma_{n-1}^2, -\eta^2), \eta\sigma_1]$. This still admits several possibilities; see Figure 6. First of all there are two critical points ($\alpha = -\sigma_n^2$ and $\alpha = -\sigma_{n-1}^2$) which could be α^o. Second, α^o could be in either interval ($(-\sigma_n^2, \eta\sigma_1)$ or the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$).

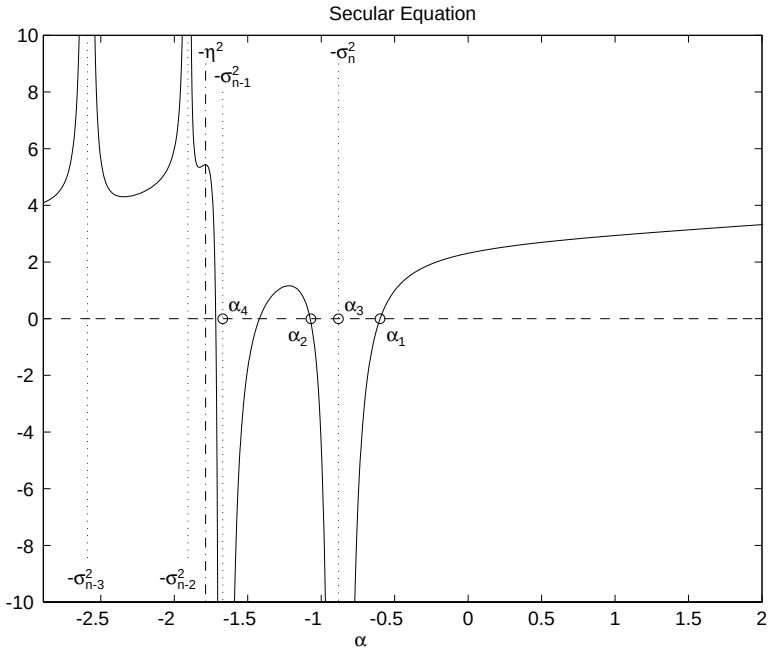

FIG. 6. *Secular equation.*

In particular, note that the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ can have multiple zeros in it, so we must also deal with this possibility. We put the arguments that only the rightmost root in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ is a candidate to be α^o in Appendix B. With this dealt with there are only four candidates zeros of $g(\alpha)$ to handle, which we denote by α_1 through α_4. We thus introduce the four candidates: $\alpha_1 \in (-\sigma_n^2, \eta\sigma_1]$, α_2 is the rightmost root in $(-\sigma_{n-1}^2, -\sigma_n^2)$, $\alpha_3 = -\sigma_n^2$, and $\alpha_4 = -\sigma_{n-1}^2$ (see Figure 6). To show that α^o is the unique root in $[-\sigma_n^2, \eta\sigma_1]$, we examine six cases. Most of the work is involved at this stage, and hence most of the mathematical difficulties occur here. The basic idea is to eliminate the possibility that any root except the one that occurs in the interval $[\max(-\sigma_n^2, -\eta^2), \eta\sigma_1]$ can be α^o. Additionally we must show the existence and uniqueness of the zero. With this established we can then use bisection or Newton's method to find the root in our algorithm.

You might be wondering why we need to use six cases to prove the assertion that α^o lies in the interval $[\max(-\sigma_n^2, -\eta^2), \eta\sigma_1]$. The reason lies in three basic factors which affect the shape of the secular equation. The first and most obvious is the size of η. Note, for instance, that if $\eta < \sigma_n$, then only one of the zeros α_1 is a candidate for α^o since we have from an earlier condition (first order condition on the Lagrangian) that $\alpha^o > -\eta^2$. Obviously to consider some of the candidates, such as α_4, we need to assume that η is large enough to admit the possibility. The cases just let us organize the assumptions into convenient groups to handle. See Figure 7. The dotted vertical lines mark where the singular values are, and the dash-dotted vertical line indicates

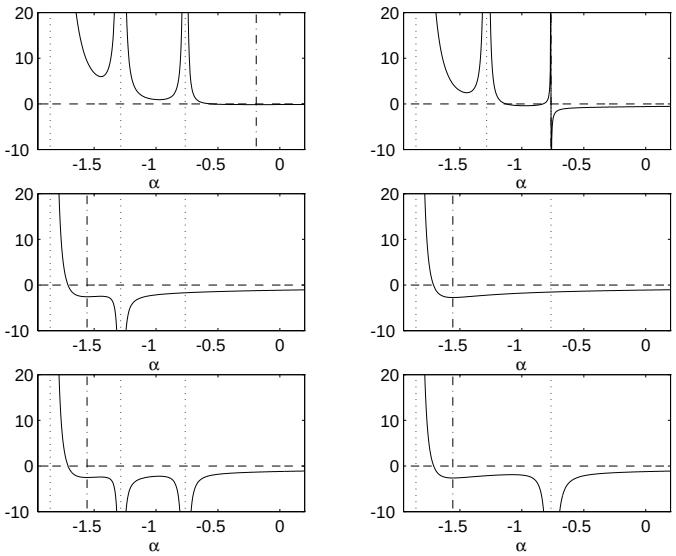

FIG. 7. *Six cases of proof. (UL) Case* 1*:* $\eta < \sigma_n$*; (UR) Case* 2*:* $\eta = \sigma_n$*; (ML) Case* 3*:* $\eta > \sigma_n$*,* b *is orthogonal to the left singular vector of* σ_n*, and* $\sigma_n < \sigma_{n-1}$*; (MR) Case* 4*:* $\eta > \sigma_n$*,* b *is orthogonal to the left singular vectors of* σ_n*, and* σ_n *has multiplicity* k*; (LL) Case* 5*:* $\eta > \sigma_n$*,* b *is not orthogonal to the left singular vector of* σ_n*, and* $\sigma_n < \sigma_{n-1}$*; (LR) Case* 6*:* $\eta > \sigma_n$*,* b *is not orthogonal to the left singular vectors of* σ_n*, and* σ_n *has multiplicity* k*.*

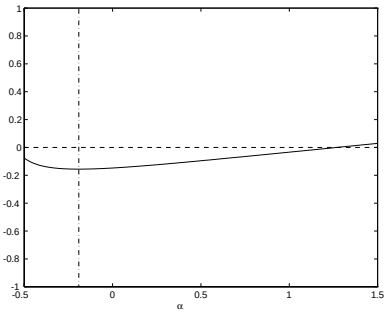

FIG. 8. *Expanded view of Case* 1 *zero.*

where $-\eta^2$ is. We note that in Case 1 of Figure 7, it looks like the secular equation becomes flat to the right of $\alpha = -0.5$ but it does not. The scale makes the graph hard to read, so we provide an expanded view of the region in Figure 8. In Case 1 we consider η small ($\eta < \sigma_n$), in Case 2 we consider the special case of $\eta = \sigma_n$, and finally in Cases 3–6 we consider η to be large ($\eta > \sigma_n$).

When η is large there are more possibilities. The first is that the smallest singular value might have multiplicity of two or more. This can be exploited to simplify the problem. In particular, α_2 does not exist in this case, and $\alpha_3 = \alpha_4$. The cases where $\sigma_n < \sigma_{n-1}$ are the more difficult ones. The second is that b might be orthogonal to the left singular vector(s) of A, which correspond to smallest singular value. This drastically changes the shape of the graph of the secular equation in the region around $\alpha = -\sigma_n^2$. See, for instance, the middle left graph in Figure 7. The pole which normally appears at $-\sigma_n^2$ is not present. In fact, the only time α_3 can be α^o is when

b is orthogonal to the left singular vector(s) of A, which corresponds to the smallest singular value (σ_n). Similarly, the only time α_4 can be α^o is when b is orthogonal to the left singular vector(s) of A, which corresponds to the second smallest singular value (σ_{n-1}). Note that if the smallest singular value has multiplicity of at least two, then $\sigma_n = \sigma_{n-1}$. This case is shown on the middle right graph of Figure 7. The last four cases cover all the combinations of singular value multiplicity and b vector orthogonality which occurs when η is large.

5. Algorithm. For the reader's convenience we present pseudocode for the algorithm in this section. The syntax has been designed to be Matlab-like. Three lines deserve particular attention, though. The first one to appear states "solve nondegenerate problem." In this case the problem is not degenerate so you will need to provide code for the nondegenerate case as outlined in [2]. The next line that could be confusing starts with "pick any Θ." In this case any unit vector, Θ, will solve the problem. An additional condition could be placed on the solution, $\hat{x}$, to select a specific Θ or to meet special requirements of the specific problem, so we leave it unspecified in our pseudocode. The final line that requires clarification starts with $\alpha \in \left[\max(-\sigma_n^2, -\eta^2), \eta\sigma_1\right]$. In this case you are finding the root of $g(\alpha)$ in the specified range, so any root finder you prefer (for instance, bisection or Newton's method) can be used.

$[U, \Sigma, V] = SVD(A)$;
$b_1 = U^T b$;
$cond = 0$;
if ($\eta < \sigma_n$) or ($\eta = \sigma_n$ and $b_1(n) = 0$)
  if ($b^T(I - A(A^TA - \eta^2 I)^{-1}A^T)b > 0$)
    solve nondegenerate problem
  else
    $cond = 1$;
  end
else
  if ($\eta = \sigma_n$)
    $cond = 1$;
  else
    if ($\sigma_n < \sigma_{n-1}$) and ($b_1(n) = 0$) and ($g(-\sigma_n^2) \geq 0$)
      $\bar{\Sigma}_1 = \Sigma(1 : n-1, 1 : n-1)$;
      $\bar{b}_1 = b_1(1 : n-1)$;
      $\hat{x} = V \begin{bmatrix} (\bar{\Sigma}_1^2 - \sigma_n^2 I)^{-1}\bar{\Sigma}_1\bar{b}_1 \\ \pm\sqrt{\frac{g(-\sigma_n^2)}{\eta^2 - \sigma_n^2}} \end{bmatrix}$;
    elseif ($\sigma_n = \sigma_{n-k+1} < \sigma_{n-k}$) and ($\|b_1(n-k+1 : n)\| = 0$)
        and ($g(-\sigma_n^2) \geq 0$)
      $\bar{\Sigma}_1 = \Sigma(1 : n-k+1, 1 : n-k+1)$;
      $\bar{b}_1 = b_1(1 : n-k+1)$;
      $r = \sqrt{\frac{g(-\sigma_n^2)}{\eta^2 - \sigma_n^2}}$;
      Pick any $\Theta \in \mathbf{R}^k$ such that $\|\Theta\| = 1$;
      $\hat{x} = V \begin{bmatrix} (\bar{\Sigma}_1^2 - \sigma_n^2 I)^{-1}\bar{\Sigma}_1\bar{b}_1 \\ r\Theta \end{bmatrix}$;
    else
      $cond = 1$;
    end

```
    end
  end
  if cond == 1
    α ∈ [max(−σ_n^2, −η^2), ησ_1] such that g(α) = 0
    x̂ = (A^T A + αI)^† A^T b;
  end
```

Where $g(\alpha)$ is given by

$$g(\alpha) = b_2^T b_2 + b_1^T \left(\Sigma_1^2 + \alpha I\right)^{-2} \left(\alpha^2 I - \eta^2 \Sigma_1^2\right) b_1,$$

and

$$\begin{aligned} A &= \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} V^T, \\ b_1 &= U_1^T b, \\ b_2 &= U_2^T b. \end{aligned}$$

6. Minimization over E. We now start the full proof. The proof will be much easier to read if the reader is familiar with the preceding papers [1, 2]. For the reader's convenience we will place major milestones in the proof in boxes at the end of the sections where the milestone occurs. We assume that the problem is degenerate and in particular that there exists an x such that $\eta\|x\| \geq \|Ax - b\|$. We will soon provide equivalent computable criteria for degeneracy; however, this formulation is more useful for the present. Our goal in this section is to reduce the problem to an equivalent formulation that does not involve E. The goal is accomplished by showing the degenerate problem is equivalent to requiring the solution to be in the set $\{x|\eta\|x\| \geq \|Ax - b\|\}$. We begin by showing that the problem requires that we be in the set, then show that any $\hat{x}$ in the set solves the problem. Note that the method used to get E is related to the formulation in [5], though we provide the full argument for the ease of the reader. Under the assumption that the problem is degenerate it follows that

$$\min_x \min_{\|E\| \leq \eta} \|Ax - b + Ex\| = 0,$$

since for any x such that $\eta\|x\| \geq \|Ax - b\|$ we can choose

$$E = -\gamma\eta \frac{(Ax - b)x^T}{\|Ax - b\|\|x\|}, \quad 0 \leq \gamma \leq 1,$$

and obtain

$$0 \leq \min_x \min_{\|E\| \leq \eta} \|Ax - b + Ex\| \leq \|Ax - b\| \left|1 - \gamma \frac{\eta\|x\|}{\|Ax - b\|}\right|,$$

and for the choice

$$\gamma = \frac{\|Ax - b\|}{\eta\|x\|} \leq 1,$$

the upper bound is zero. Since there exists an E which makes the minimum zero, the minimum value of the norm is zero. Therefore we only need consider the equation

$$Ax - b + Ex = 0 \tag{6.1}$$

with the constraint $\|E\| \leq \eta$. This constrained equation is equivalent to being on the set defined by

$$\|Ax - b\| \leq \eta \|x\|. \tag{6.2}$$

To prove this, we first show that if the constrained equation (6.1) is met, then we are in the set (6.2).

$$Ax - b + Ex = 0,$$

$$Ax - b = -Ex.$$

Taking the norm of both sides we obtain

$$\|Ax - b\| = \|Ex\|$$

and we note that this implies

$$\|Ax - b\| \leq \|E\| \|x\|.$$

Then using the constraint on the perturbation size, $\|E\| \leq \eta$, we obtain

$$\|Ax - b\| \leq \eta \|x\|$$

and we have the desired result. We now show that if we are in the set (6.2), then the constraint equation (6.1) is met. This is accomplished by showing that for any x in the set, there exists a perturbation, E_0, such that the constraint equation is satisfied. To do this consider

$$E_0 = -\frac{(Ax - b)x^T}{\|x\|^2}.$$

We first note that this perturbation satisfies the constraint on the size of the perturbations ($\|E\| \leq \eta$).

$$\|E_0\| \leq \frac{\|Ax - b\|}{\|x\|}.$$

Since on the set $\|Ax - b\| \leq \eta \|x\|$ we have

$$\|E_0\| \leq \eta,$$

we now consider the equation given by $Ax - b + E_0 x$. We note that this is

$$Ax - b + E_0 x = Ax - b - (Ax - b).$$

Thus we have trivially that $Ax - b + E_0 = 0$ and the assertion is proven.

We know there are multiple solutions which will solve the problem as stated. Since any will solve the original problem, we are free to add an additional constraint which will simplify the solution and ensure the solution meets other requirements. A reasonable choice is to pick the solution with the minimum norm. Other nice properties of this choice also recommend it. For instance, it is possible under certain conditions for the min-max solution (from [1]) to also solve the degenerate min-min

problem. When this occurs the min-max solution is the solution to the degenerate problem with minimum norm. We do not prove this for reasons of space, but it does provide a good understanding of the relationships between the problems and gives additional motivation for the choice. Using the choice of the minimum norm solution, the problem can be rewritten into the better form, as follows.

The degenerate problem can be reformulated as a unique problem by considering

$$\min_{\|Ax-b\| \leq \eta \|x\|} \|x\|.$$

7. Computable conditions for degeneracy. The constraint, $\|Ax - b\| \leq \eta\|x\|$, defines the set on which our solution lies and is thus referred to as the feasibility constraint. The feasibility constraint can be squared and expanded to obtain

$$x^T A^T A x - 2x^T A^T b + b^T b \leq \eta^2 x^T x. \tag{7.1}$$

Let $A = U\Sigma V^T$ be the SVD of A conformally partitioned as follows:

$$U = \begin{pmatrix} U_1 & U_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_1 \\ 0 \end{pmatrix},$$

and define both $b_i = U_i^T b$ for $i = 1, 2$, and $z = V^T x$. These definitions are made solely to simplify the expressions we are working with and provide a convenient shorthand for the rest of the problem. Then inequality (7.1) can be simplified to obtain

$$z^T \Sigma_1^2 z - 2z^T \Sigma_1 b_1 + b_1^T b_1 + b_2^T b_2 \leq \eta^2 z^T z. \tag{7.2}$$

Now assuming that the singular values are in decreasing order, partition Σ_1 as follows:

$$\Sigma_1 = \begin{pmatrix} \Sigma_+ & 0 \\ 0 & \Sigma_- \end{pmatrix},$$

where $\Sigma_+^2 - \eta^2 I \geq 0$ and $\Sigma_-^2 - \eta^2 I < 0$. Also conformally partition z and b_1

$$z = \begin{pmatrix} z_+ \\ z_- \end{pmatrix} \quad b_1 = \begin{pmatrix} b_{1+} \\ b_{1-} \end{pmatrix}.$$

Then inequality (7.2) can be expanded into

$$\begin{aligned} 0 \geq & \\ & z_+^T(\Sigma_+^2 - \eta^2 I)z_+ - 2z_+^T \Sigma_+ b_{1+} + b_{1+}^T b_{1+} \\ & + z_-^T(\Sigma_-^2 - \eta^2 I)z_- - 2z_-^T \Sigma_- b_{1-} + b_{1-}^T b_{1-} \\ & + b_2^T b_2. \end{aligned}$$

Now we observe that if Σ_- is nonempty, then the inequality always has at least one z which makes it true. In other words if $A^T A - \eta^2 I$ is indefinite, then the problem is always degenerate. On the other hand, if $A^T A - \eta^2 I$ is positive-semidefinite, then degeneracy depends on the vector b. To get a computable condition for degeneracy, we first note that when $x = 0$ we have that the constraint is nonnegative. We proceed by minimizing the expression

$$x^T(A^T A - \eta^2 I)x - 2x^T A^T b + b^T b$$

and when $\eta \neq \sigma_i$ we obtain

$$x_o = (A^T A - \eta^2 I)^{-1} A^T b.$$

Now, when $A^T A - \eta^2 I$ is positive, we must have that the constraint is nonpositive at this point. On plugging this back into the expression being minimized we obtain

$$b^T (I - A(A^T A - \eta^2 I)^{-1} A^T) b \leq 0 \tag{7.3}$$

as the required computable condition for the problem to be degenerate when $\eta < \sigma_n$.

The problem is degenerate if either

$$\eta > \sigma_n$$

or

$$b^T (I - A(A^T A - \eta^2 I)^{-1} A^T) b \leq 0.$$

8. Solution is on the boundary. We want to establish that the optimal solution is obtained at the boundary of the feasible set; that is, at the minimum norm solution the inequality is actually an equality. Mathematically this means the feasibility constraint, $\|Ax - b\| \leq \eta \|x\|$, is actually an equality, $\|Ax - b\| = \eta \|x\|$. To prove this we use the shorthand developed in the last section that given the SVD of A, then $b_i = U_i^T b$ for $i = 1, 2$, and $z = V^T x$. The problem of finding the solution with the smallest norm to the degenerate problem can now be recast as minimizing $z^T z$ subject to the inequality constraint (7.3).

Now if $b = 0$, then clearly the minimum norm solution is $z = 0$ which does lie on the boundary ($0 = 0$). So we restrict ourselves to the case when $b \neq 0$. Let us denote by $f(z)$ the expression on the left-hand side of inequality (7.3). Then it is clear that $f(0) > 0$, and therefore $z = 0$ is not a feasible point. Now suppose that contrary to our hypothesis that the optimal solution occurs at an interior point. Denote that optimal solution by z_0. Since it is an interior point we must have $0 > f(z_0)$. Let γ denote a scalar and consider the function $f(\gamma z_0)$ as γ varies. Since $f(\cdot)$ is a continuous function it follows that as γ is decreased from 1 towards 0, the value of $f(\gamma z_0)$ must at sometime become equal to 0. But now we have a contradiction as $\|\gamma z_0\| < \|z_0\|$ for $0 < \gamma < 1$. Hence we prove our hypothesis that the optimal solution must lie on the boundary of the feasible set.

Therefore we can restrict our attention to the problem

$$\min_{\|Ax-b\| = \eta \|x\|} \|x\|.$$

We note that the problem is unaffected by squaring, thus to simplify the algebra we will work with the squared problem.

The problem is equivalently stated as

$$\min_{\|Ax-b\|^2 = \eta^2 \|x\|^2} \|x\|^2.$$

9. Reduction to secular equation. Since we have reduced the problem to an equality constrained minimization problem, we can use the method of Lagrange multipliers. Letting λ denote the Lagrange multiplier we obtain the following set of equations that characterize the critical points

$$x + \lambda \left(A^T (Ax - b) - \eta^2 x \right) = 0.$$

Simplifying, we obtain

$$\left(A^TA + \frac{1-\lambda\eta^2}{\lambda} I\right) x = A^Tb.$$

Make the definition $(1-\lambda\eta^2)/\lambda = \alpha$. Then we have

$$x = (A^TA + \alpha I)^{-1}A^Tb.$$

Plugging this into $\|Ax-b\|^2 = \eta^2\|x\|^2$ and using the SVD of A we obtain

$$b_2^Tb_2 + b_1^T\Sigma_1^4(\Sigma_1^2+\alpha I)^{-2}b_1 - 2b_1^T\Sigma_1^2(\Sigma_1^2+\alpha I)^{-1}b_1 + b_1^Tb_1 = \eta^2 b_1^T\Sigma_1^2(\Sigma_1^2+\alpha I)^{-2}b_1.$$

Simplifying we get

$$b_2^Tb_2 + b_1^T(\Sigma_1^2+\alpha I)^{-2}(\alpha^2 I - \eta^2\Sigma_1^2)b_1 = 0.$$

Since we are interested in finding the values of α for which the right-hand side of the above equation is zero, we define the function $g(\alpha)$ as

$$g(\alpha) = b_2^Tb_2 + b_1^T(\Sigma_1^2+\alpha I)^{-2}(\alpha^2 I - \eta^2\Sigma_1^2)b_1$$

and then study the zeros of this function. The function $g(\alpha)$ is called the "secular equation," since it is rational function of one variable. If σ_i denotes the ith singular value of A, then the above secular equation has poles at $-\sigma_i^2$.

This secular equation can have up to $2n$ real zeros. One of them will give us the minimum norm solution to our problem, $x(\alpha^o)$. We note that if $\alpha > \eta\sigma_1$ in the secular equation, then we must have $b = 0$, which as we stated earlier requires $z = 0$, and thus $x = 0$. Since we are considering $b \neq 0$ we must have $\alpha \leq \eta\sigma_1$.

The secular equation, $g(\alpha)$ is given by

$$g(\alpha) = b_2^Tb_2 + b_1^T(\Sigma_1^2+\alpha I)^{-2}(\alpha^2 I - \eta^2\Sigma_1^2)b_1.$$

10. Main theorem. We claim that in all cases where a degenerate solution exists, the minimum norm solution is determined by the unique root of the secular equation in the interval $[\max(-\sigma_n^2, -\eta^2), \eta\sigma_1]$.

The rest of the paper is devoted to establishing this claim. This is a difficult task due to the nonconvex nature of the problem and the presence of multiple local minima.

The solution to the problem, $\hat{x}$, is given by $\hat{x} = x(\alpha^o)$ with α^o the unique zero of

$$g(\alpha) = b_2^Tb_2 + b_1^T(\Sigma_1^2+\alpha I)^{-2}(\alpha^2 I - \eta^2\Sigma_1^2)b_1$$

in the interval $[\max(-\sigma_n^2, -\eta^2), \eta\sigma_1]$.

11. First and second order conditions. Since the Lagrange multiplier must be nonnegative at a local minimum and $\lambda = 1/(\alpha+\eta^2)$ we conclude that

$$\alpha \geq -\eta^2. \tag{11.1}$$

To narrow down the interesting zeros we look at the second order conditions for a local minimum. Our Lagrangian was

$$L(x,\lambda) = \|x\|^2 + \lambda(\|Ax-b\|^2 - \eta^2\|x\|^2).$$

The second order condition for a local minimum is that the Hessian of $L(x, \lambda)$ with respect to x be positive-semidefinite when restricted to the tangent subspace of the constraint. Differentiating once we have

$$\nabla_x L(x, \lambda) = 2x + \lambda \left(2A^T(Ax - b) - 2\eta^2 x\right).$$

Differentiating once more we get

$$\nabla_x^2 L(x, \lambda) = 2I + \lambda \left(2A^T A - 2\eta^2 I\right),$$

which on simplifying yields

$$\nabla_x^2 L(x, \lambda) = 2\lambda \left(\alpha I + A^T A\right).$$

The constraint is

$$c(x) = \|Ax - b\|^2 - \eta^2 \|x\|^2.$$

The gradient of the constraint is

$$\nabla_x c(x) = 2A^T(Ax - b) - 2\eta^2 x,$$

which can be simplified by noting that

$$A^T(Ax - b) = -\alpha x$$

thus

$$\nabla_x c(x) = -(\alpha + \eta^2)x.$$

The tangent subspace of the constraint has $n - 1$ dimensions (even when $\eta = \sigma_i$). We now construct a basis for this subspace. Using the SVD notation developed in section 7 we have

$$V^T \nabla_x c(x) = -(\alpha + \eta^2)z.$$

Similarly we can change the basis for the Hessian of the Lagrangian

$$V^T \nabla_x^2 L(x, \lambda) V = 2\lambda \left(\Sigma_1^2 + \alpha I\right).$$

We partition z as

$$z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix},$$

where z_1 is a scalar. Let

$$H = \begin{pmatrix} z_2^T \\ -z_1 I \end{pmatrix}.$$

Then $H^T z = 0$. Therefore the restricted Hessian is

$$H^T V^T \nabla_x^2 L(x, \lambda) V H = 2\lambda \left(H^T \Sigma_1^2 H + \alpha H^T H\right).$$

We note that the second order condition requires that the restricted Hessian be positive-semidefinite, and so we can apply Cauchy's interlacing theorem. Cauchy's

interlacing theorem tells us that the smallest eigenvalue for this matrix must lie between the smallest and second smallest eigenvalues for the nonrestricted Hessian. Thus for a local minimum the second smallest eigenvalue of the nonrestricted Hessian must be greater than zero. For the condition on the second smallest eigenvalue to be met, α must satisfy the constraint $\alpha \geq -\sigma_{n-1}^2$, where σ_{n-1} is the second smallest singular value of A.

This raises the question of how many zeros of the secular equation are larger than $\max(-\eta^2, -\sigma_{n-1}^2)$ and which of them corresponds to the global minimum. We proceed by systematically eliminating zeros in this range. We have two critical points (where the secular equation becomes infinite) which correspond to $\alpha = -\sigma_{n-1}^2$ and $\alpha = -\sigma_n^2$. We also have two intervals to worry about, namely, $(-\sigma_n^2, \eta\sigma_1)$ and $(-\sigma_{n-1}^2, -\sigma_n^2)$. In the first interval we can show that there is only one zero, but this is not true for the second interval. In section 12 we use the second order condition to rule out half of the zeros in the second interval. We show in Appendix B that only the rightmost root in the second interval is actually a candidate. We are left with four candidates, two in the intervals and two critical points, and we then use six cases to prove which one corresponds to the global minimum.

$$\alpha^o > \max(-\eta^2, -\sigma_{n-1}^2).$$

12. Squeezing the second order conditions. We can use the second order conditions to discard some zeros in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$. Recall that the restricted Hessian is

$$H^T V^T \nabla_x^2 L(x,\lambda) V H = 2\lambda \left(H^T \Sigma_1^2 H + \alpha H^T H \right).$$

This can be expanded to obtain

$$H^T V^T \nabla_x^2 L(x,\lambda) V H = 2\lambda \left(\sigma_1^2 z_2 z_2^T + z_1^2 \Sigma_2^2 + \alpha z_2 z_2^T + \alpha z_1^2 I \right),$$

where

$$\Sigma_1 = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}.$$

We then make the conformal partition

$$b_1 = \begin{pmatrix} b_{11} \\ b_{12} \end{pmatrix}$$

and use the representation $z = (\Sigma_1^2 + \alpha I)^{-1} \Sigma_1 b_1$ to simplify the expansion. Additionally, we make the definition $M = H^T V^T \nabla_x^2 L(x,\lambda) V H$ for ease of reading and we can thus get the simplified expansion

$$M = 2\lambda \frac{b_{11}^2 \sigma_1^2}{(\sigma_1^2 + \alpha)^2} (\Sigma_2^2 + \alpha I) \left(I + \frac{(\sigma_1^2 + \alpha)^3}{b_{11}^2 \sigma_1^2} (\Sigma_2^2 + \alpha I)^{-2} \Sigma_2 b_{12} b_{12}^T \Sigma_2 (\Sigma_2^2 + \alpha I)^{-1} \right).$$

We now compute the determinant,

$$\det(M) = \left(\frac{2\lambda b_{11}^2 \sigma_1^2}{(\sigma_1^2 + \alpha)^2} \right)^n \det(\Sigma_2^2 + \alpha I) \left(1 + \frac{(\sigma_1^2 + \alpha)^3}{b_{11}^2 \sigma_1^2} b_{12}^T \Sigma_2^2 (\Sigma_2^2 + \alpha I)^{-3} b_{12} \right),$$

which can be further simplified to obtain

$$\det(M) = \left(\frac{2\lambda b_{11}^2\sigma_1^2}{(\sigma_1^2+\alpha)^2}\right)^n \frac{(\sigma_1^2+\alpha)^3}{b_{11}^2\sigma_1^2}\det(\Sigma_2^2+\alpha I)\left(b_1^T\Sigma_1^2(\Sigma_1^2+\alpha I)^{-3}b_1\right). \tag{12.1}$$

We recall the definition of the secular equation, $g(\alpha)$, given in section 9:

$$g(\alpha) = b_2^Tb_2 + b_1^T(\Sigma_1^2+\alpha I)^{-2}(\alpha^2 I - \eta^2\Sigma_1^2)b_1.$$

Then differentiating once we obtain

$$g'(\alpha) = 2(\alpha+\eta^2)b_1^T\Sigma_1^2(\Sigma_1^2+\alpha I)^{-3}b_1. \tag{12.2}$$

Using this we can rewrite (12.1) as

$$\det(M) = \left(\frac{2\lambda b_{11}^2\sigma_1^2}{(\sigma_1^2+\alpha)^2}\right)^n \frac{(\sigma_1^2+\alpha)^3}{2(\alpha+\eta^2)b_{11}^2\sigma_1^2}\det(\Sigma_2^2+\alpha I)g'(\alpha).$$

Therefore we see that when a root of the secular equation lies in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$, then it can correspond to a local minimum only if $g'(\alpha)$ is nonpositive. **This essentially means that only half of the zeros in the interval correspond to local minima.**

A zero, α_k, of $g(\alpha)$ in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ can correspond to a local minimum of the Lagrangian (and thus have a chance of being the global minimum α^o) only if

$$g'(\alpha_k) \le 0.$$

13. Four candidate zeros. At this point we can see several potential candidates for α. First, we have the possibility of a root in the interval $[-\sigma_n^2, \eta\sigma_1]$ designated α_1. The uniqueness and conditions for existence of α_1 will be shown later. Second, we potentially have many roots in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$, but only the rightmost one matters as will be shown later and it is thus designated α_2. Finally, we could have up to two critical points, $\alpha_3 = -\sigma_n^2$ and $\alpha_4 = -\sigma_{n-1}^2$. We summarize the candidates in Table 1.

TABLE 1
Candidate zeros.

$\alpha_1 \in [-\sigma_n^2, \eta\sigma_1]$
$\alpha_2 \in (-\sigma_{n-1}^2, -\sigma_n^2)$
$\alpha_3 = -\sigma_n^2$
$\alpha_4 = -\sigma_{n-1}^2$

The proof involves six cases, which cover special conditions for the problem. See Table 2. The first two cases involve small values of η. The second two cases cover

when b is orthogonal to the left singular vector(s) of the smallest singular value. The last two cases cover when b is not orthogonal to the left singular vector(s) of the smallest singular value. We now proceed to prove this and show which candidate root will yield the solution to the problem, $\hat{x}$.

TABLE 2
Six cases of the proof.

Case 1: $\eta < \sigma_n$
Case 2: $\eta = \sigma_n$
Case 3: $\eta > \sigma_n$, $b_{1,n} = 0$, $\sigma_n < \sigma_{n-1}$
Case 4: $\eta > \sigma_n$, $\|b_{1,(n-k+1,n)}\| = 0$, $\sigma_n = \sigma_{n-k+1}$
Case 5: $\eta > \sigma_n$, $b_{1,n} \neq 0$, $\sigma_n < \sigma_{n-1}$
Case 6: $\eta > \sigma_n$, $\|b_{1,n-k+1}\| \neq 0$, $\sigma_n = \sigma_{n-k+1}$

14. Case 1: $\eta < \sigma_n$. There is only one root in the interval $[-\eta^2, \eta\sigma_1]$ and this must correspond to the global minimum, as there are no other local minima to worry about. The only candidate zero is α_1 because of the first order condition, (11.1). We need to only prove the existence and uniqueness of α_1.

Since $\alpha + \eta^2 \geq 0$ from (11.1), it follows by using (12.2) that $g'(\alpha)$ is positive in the interval $(-\eta^2, \infty)$ when $\eta \leq \sigma_n$. Therefore, there can be at most one root in the interval $[-\eta^2, \eta\sigma_1]$.

We now show that there is at least one root in the interval $[-\eta^2, \eta\sigma_1]$. Simplifying the degeneracy condition in (7.3) by using the SVD of A we obtain

$$b_2^T b_2 - \eta^2 b_1^T (\Sigma_1^2 - \eta^2 I)^{-1} b_1 \leq 0,$$

which is identical to $g(-\eta^2) \leq 0$. Furthermore,

$$\lim_{\alpha \to \eta\sigma_1} g(\alpha) > 0.$$

Therefore, there must be a zero of $g(\alpha)$ in the interval $[-\eta^2, \eta\sigma_1]$.

15. Case 2: $\eta = \sigma_n$. We claim that there is a unique root of $g(\alpha)$ in $[-\sigma_n^2, \eta\sigma_1]$, and this is the global minimum. Uniqueness is established by the same method as in section 14, and thus if a root exists in the interval $[-\sigma_n^2, \eta\sigma_1]$, it is unique. Only two candidates, the zero α_1 and the critical point α_3, are possible because of the first order condition, (11.1). We will proceed to prove the claim in two steps. Before we start

the first step we note that if σ_n is multiple with multiplicity k, then $\tilde{b}_1 = b_{1,(n-k+1:n)}$ is the partitioning of b_1 corresponding to the multiple singular values of σ_n.

The first case is when $b_{1,n} \neq 0$ or $\|\tilde{b}_1\| \neq 0$. We first note that in this case the candidate zero α_3 is not possible. To see this we first partition Σ_1 as

$$\Sigma_1 = \begin{pmatrix} \bar{\Sigma}_1 & 0 \\ 0 & \sigma_n \end{pmatrix}.$$

We similarly partition z into $\bar{z}$ and z_n, and b_1 into $\bar{b}_1$ and $b_{1,n}$. We can use these to rewrite the Lagrange condition, $(A^TA + \alpha I)x = A^Tb$, as

$$\begin{pmatrix} \bar{\Sigma}_1^2 + \alpha_3 I & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \bar{z} \\ z_n \end{pmatrix} = \begin{pmatrix} \bar{\Sigma}_1 \bar{b}_1 \\ b_{1,n} \end{pmatrix}.$$

Since $b_{1,n} \neq 0$, we see that α_3 cannot be α^o. The existence of a root in the interval $(-\sigma_n^2, \eta\sigma_1)$ follows from the observation that

$$\lim_{\alpha \to -\sigma_n^2 +} g(\alpha) = -\infty,$$
$$\lim_{\alpha \to \eta\sigma_1} g(\alpha) \geq 0.$$

Thus when $b_{1,n} \neq 0$ or $\|\tilde{b}_1\| \neq 0$, $\alpha^o = \alpha_1$.

The second case is $b_{1,n} = 0$ when $\sigma_n < \sigma_{n-1}$ or $\|\tilde{b}_1\| = 0$ when σ_n is multiple. In this case we note that there is no longer a pole in $g(\alpha)$ at $\alpha = -\sigma_n^2$. By observing the degeneracy condition given by (7.3) that the degeneracy in this case is determined by b so for degeneracy, (7.3) must hold for a smaller problem. Simplifying the (7.3) using the SVD of A we obtain

$$b_2^T b_2 - \eta^2 b_1^T (\Sigma_1^2 - \eta^2 I)^{-1} b_1 \leq 0,$$

which is identical to $g(-\eta^2) \leq 0$. Furthermore,

$$\lim_{\alpha \to \eta\sigma_1} g(\alpha) \geq 0.$$

Therefore, there must be a root in the interval $[-\eta^2, \eta\sigma_1]$, so α_1 exists. We will show that when α_3 is α^o, then $\alpha_3 = \alpha_1$. To satisfy the equation

$$\begin{pmatrix} \bar{\Sigma}_1^2 + \alpha_3 I & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \bar{z} \\ z_n \end{pmatrix} = \begin{pmatrix} \bar{\Sigma}_1 \bar{b}_1 \\ 0 \end{pmatrix},$$

we must have that

$$\bar{z} = \left(\bar{\Sigma}_1^2 + \alpha_3 I\right)^{-1} \bar{\Sigma}_1 \bar{b}_1.$$

The constraint equation can be written in z and simplified to

$$\alpha_3 \bar{b}_1^T \left(\bar{\Sigma}_1^2 + \alpha_3 I\right)^{-1} \bar{b}_1 + b_2^T b_2 = 0.$$

We note that this is exactly $g(\alpha_3) = 0$. Thus for α_3 to be a candidate it must also be the unique root in the interval $[-\eta^2, \eta\sigma_1]$. The condition for α_3 to be α^o is that $\alpha_1 = \alpha_3$, and thus we can easily see that in all cases the unique zero which corresponds to the problem solution, $x(\alpha^o)$, is given by α_1.

16. Case 3: $\eta > \sigma_n$, $b_{1,n} = 0$, $\sigma_n < \sigma_{n-1}$. We claim that there is a unique root in $[-\sigma_n^2, \eta\sigma_1]$ and this is the global minimum. We now establish this claim. Two cases arise when $b_{1,n} = 0$ by observing the equation

$$\begin{pmatrix} \bar{\Sigma}_1^2 + \alpha I & 0 \\ 0 & \sigma_n^2 + \alpha \end{pmatrix} \begin{pmatrix} \bar{z} \\ z_n \end{pmatrix} = \begin{pmatrix} \bar{\Sigma}_1 \bar{b}_1 \\ 0 \end{pmatrix}. \tag{16.1}$$

First, we could have $\alpha = \alpha_3 = -\sigma_n^2$, which we note can only happen when $b_{1,n} = 0$. The second case is $z_n = 0$. First, we note that we still have

$$\lim_{\alpha \to \eta\sigma_1} g(\alpha) \geq 0.$$

We also know that $g'(\alpha) > 0$ on the interval $(-\sigma_{n-1}^2, \infty)$, thus if a root exists, it is unique. We will start by finding the form of the solution $\hat{x}$ when $\alpha = \alpha_3$ and then we will show the conditions for determining which candidate zero yields the global minimum.

When $\alpha = -\sigma_n^2$ the solution is found in two steps. First we solve for $\bar{z}$ from (16.1). We obtain

$$\bar{z} = \left(\bar{\Sigma}_1^2 - \sigma_n^2 I\right)^{-1} \bar{\Sigma}_1 \bar{b}_1.$$

We note that the constraint can be written in z as

$$\left\| \begin{matrix} \Sigma_1 z - b_1 \\ b_2 \end{matrix} \right\|^2 - \eta^2 \|z\|^2 = 0.$$

We now separate z_n in the constraint and obtain

$$\bar{b}_1^T \left(\bar{\Sigma}_1^2 - \sigma_n^2 I\right)^{-2} \left(\sigma_n^4 I - \eta^2 \bar{\Sigma}_1^2\right) \bar{b}_1 + b_2^T b_2 + \left(\sigma_n^2 - \eta^2\right) z_n^2 = 0.$$

We note that this can be rewritten in terms of $g(-\sigma_n^2)$ as

$$g\left(-\sigma_n^2\right) + \left(\sigma_n^2 - \eta^2\right) z_n^2 = 0.$$

We thus see there are two answers (positive and negative squares) for z_n. The answers for z_n are given by

$$z_n^2 = \frac{g\left(-\sigma_n^2\right)}{\eta^2 - \sigma_n^2}. \tag{16.2}$$

Note that for a solution for z_n to exist we must have $g(-\sigma_n^2) \geq 0$. The solution is then given by

$$\hat{x} = V \begin{bmatrix} \left(\bar{\Sigma}_1^2 - \sigma_n^2 I\right)^{-1} \bar{\Sigma}_1 \bar{b}_1 \\ \pm\sqrt{\frac{g(-\sigma_n^2)}{\eta^2 - \sigma_n^2}} \end{bmatrix}.$$

We still need to identify which of the potential roots is the actual one we want. We break this into two steps. The first is when $g(-\sigma_n^2) \leq 0$, and the second is $g(-\sigma_n^2) > 0$. If $g(-\sigma_n^2) \leq 0$, then we trivially have a unique root in $[-\sigma_n^2, \eta\sigma_1]$. Moreover, no root exists in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ so α_2 is not a candidate. We note that for $\alpha_4 = -\sigma_{n-1}^2$ to be a candidate, it must be true that $b_{1,n-1} = 0$. When $b_{1,n-1} = 0$, we have $g'(\alpha) > 0$ on the interval $(-\sigma_{n-2}^2, \infty)$, which means $g(-\sigma_{n-1}^2) < 0$. If we assume

$\alpha = \alpha_4$ and proceed similarly to section 16 we see that we must have $g(-\sigma_{n-1}^2) \geq 0$ and thus α_4 cannot be α^o. Note that when $g(-\sigma_n^2) < 0$, it is impossible for $\alpha = -\sigma_n^2$. When $g(-\sigma_n^2) = 0$, the unique root is $\alpha = -\sigma_n^2$ and thus the two remaining candidate zeros can easily be seen to coincide. Thus when $g(-\sigma_n^2) \leq 0$, the unique zero is given by α_1.

When $g(-\sigma_n^2) > 0$ no root exists in $(-\sigma_n^2, \eta\sigma_1]$ so α_1 is not α^o but as we saw in section 16 this is the condition for $\alpha = \alpha_3 = -\sigma_n^2$. We note that when $g(-\sigma_n^2) > 0$, there can be a root in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$, but we know that the slope is positive in this interval and by the results of section 12 it cannot be a minimum. The only remaining question in this case is if $\alpha_4 = -\sigma_{n-1}^2$ is a candidate when $g(-\sigma_n^2) > 0$. We again recall that for $-\sigma_{n-1}^2$ to be a candidate, it must be true that $b_{1,n-1} = 0$ and $g(-\sigma_{n-1}^2) \geq 0$. We must thus satisfy the equation

$$\begin{pmatrix} \tilde{\Sigma}_1^2 + \alpha I & 0 & 0 \\ 0 & \sigma_{n-1}^2 + \alpha & 0 \\ 0 & 0 & \sigma_n^2 + \alpha \end{pmatrix} \begin{pmatrix} \tilde{z} \\ z_{n-1} \\ z_n \end{pmatrix} = \begin{pmatrix} \tilde{\Sigma}_1 \tilde{b}_1 \\ 0 \\ 0 \end{pmatrix}. \tag{16.3}$$

We proceed to show that $-\sigma_{n-1}^2$ is not a candidate when $g(-\sigma_{n-1}^2) \geq 0$. We note that since $b_{1,n-1} = 0 = b_{1,n}$ we must have $g'(\alpha) > 0$ on the interval $(-\sigma_{n-2}^2, \infty)$. Now introduce the parameter $\gamma = \|b_{1,n-1}\|^2$ and we will consider a continuity argument on γ similar to the continuity argument we will consider in section 18. Since the argument is very similar to the one we will be constructing, we will only sketch the details here. Note that for $\gamma \neq 0$ we have a root in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ which is not the global minimum. As γ goes to zero we make this root move to the left, and it reaches $-\sigma_{n-1}^2$ when $\gamma = 0$, since $g(-\sigma_{n-1}^2) \geq 0$. The derivative of the cost with respect to γ can be seen to be negative in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ by the following method. First, take the derivative and note that there appears the term $d\alpha(\gamma)/d\gamma$, which we solve for by taking the derivative of $g(\alpha(\gamma)) = 0$ with respect to γ. Substituting back in and simplifying we see that as γ increases, the cost decreases in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ and thus the x corresponding to the root which appears in the interval when $\gamma \neq 0$ has a lower cost than the x which corresponds to $-\sigma_{n-1}^2$. The root is not a global minimum, however, and so neither can be α^o at $-\sigma_{n-1}^2$. The only possibility when $g(-\sigma_n^2) \geq 0$ is thus $\alpha^o = -\sigma_n^2$.

17. Case 4: $\eta > \sigma_n$, $\|b_{1,(n-k+1,n)}\| = 0$, $\sigma_n = \sigma_{n-k+1}$. We claim that there is a unique root in $[-\sigma_n^2, \eta\sigma_1]$, and this is the global minimum. We now establish this claim. For simplicity partition Σ_1 as

$$\Sigma_1 = \begin{pmatrix} \bar{\Sigma}_1 & 0 \\ 0 & \sigma_n I \end{pmatrix},$$

where $\bar{\Sigma}_1$ corresponds to the singular values that are strictly greater than σ_n. We similarly partition z into $\bar{z}$ and $\tilde{z}$, and b_1 into $\bar{b}_1$ and $\tilde{b}_1$. Two cases arise when $\tilde{b}_1 = 0$ by observing the equation

$$\begin{pmatrix} \bar{\Sigma}_1^2 + \alpha I & 0 \\ 0 & (\sigma_n^2 + \alpha) I \end{pmatrix} \begin{pmatrix} \bar{z} \\ \tilde{z} \end{pmatrix} = \begin{pmatrix} \bar{\Sigma}_1 \bar{b}_1 \\ 0 \end{pmatrix}. \tag{17.1}$$

First we could have $\alpha = -\sigma_n^2$, which we note can only happen when $b_{1,n} = 0$. The second case is $\tilde{z} = 0$. First we note that we still have

$$\lim_{\alpha \to \eta\sigma_1} g(\alpha) \geq 0.$$

We also know that $g'(\alpha) > 0$ on the interval $(-\sigma_{n-1}^2, \infty)$, thus if a root exists, it is unique.

When $\alpha = -\sigma_n^2$, the solution is found in two steps. First we solve for $\bar{z}$ from (17.1). We obtain

$$\bar{z} = \left(\bar{\Sigma}_1^2 - \sigma_n^2 I\right)^{-1} \bar{\Sigma}_1 \bar{b}_1.$$

We note that the constraint can be written in z as

$$\left\| \begin{matrix} \Sigma_1 z - b_1 \\ b_2 \end{matrix} \right\|^2 - \eta^2 \|z\|^2 = 0.$$

We now separate $\tilde{z}$ in the constraint and obtain

$$\bar{b}_1^T \left(\bar{\Sigma}_1^2 - \sigma_n^2 I\right)^{-2} \left(\sigma_n^4 I - \eta^2 \bar{\Sigma}_1^2\right) \bar{b}_1 + b_2^T b_2 + \left(\sigma_n^2 - \eta^2\right) \tilde{z}^T \tilde{z} = 0.$$

Similar to what we saw in the last section, we note that the above equation can be written in terms of $g(-\sigma_n^2)$. Doing so, we obtain

$$g\left(-\sigma_n^2\right) + \left(\sigma_n^2 - \eta^2\right) \tilde{z}^T \tilde{z} = 0.$$

We note that this defines a hypersphere with radius

$$r = \sqrt{\frac{g\left(-\sigma_n^2\right)}{\eta^2 - \sigma_n^2}}.$$

To be able to solve for the radius we must have $g(-\sigma_n^2) \geq 0$, and thus this is a condition on the solution when $\alpha = -\sigma_n^2$. Let Θ be any vector with unit Euclidean norm. The solutions for $\tilde{z}$ are given by

$$\tilde{z} = r\Theta.$$

The solution is then given by

$$\hat{x} = V \begin{bmatrix} \left(\bar{\Sigma}_1^2 - \sigma_n^2 I\right)^{-1} \bar{\Sigma}_1 \bar{b}_1 \\ r\Theta \end{bmatrix}.$$

We note that the second order condition requires that $\alpha \leq -\sigma_n^2$ and thus the only candidates are α_1 and α_3. If $g(-\sigma_n^2) \leq 0$, then we trivially have a unique root in $[-\sigma_n^2, \eta\sigma_1]$, and it is impossible for $\alpha = -\sigma_n^2$. If $g(-\sigma_n^2) > 0$, no root exists in $(-\sigma_n^2, \eta\sigma_1]$ but as we saw above this is the condition for $\alpha = -\sigma_n^2$. When $g(-\sigma_n^2) = 0$ the two zeros can easily be seen to coincide.

18. Case 5: $\eta > \sigma_n$, $b_{1,n} \neq 0$, $\sigma_n < \sigma_{n-1}$. We now claim that there is a unique root in $(-\sigma_n^2, \eta\sigma_1]$ and it is the global minimum. We note first that since $b_{1,n} \neq 0$, we cannot have $\alpha = -\sigma_n^2$.

The existence of a root in the interval $[-\sigma_n^2, \eta\sigma_1]$ follows from the observation that

$$\lim_{\alpha \to -\sigma_n^2 +} g(\alpha) = -\infty,$$
$$\lim_{\alpha \to \eta\sigma_1} g(\alpha) \geq 0.$$

Uniqueness is established by the same method as in section 14.

We now proceed to show that of the three candidate roots only the one in the interval $(-\sigma_n^2, \eta\sigma_1]$ can be the global minimum. The argument proceeds by continuation on $\beta = b_{1,n}^2$. We begin by defining

$$\bar{g}(\alpha) = b_2^T b_2 + \bar{b}_1^T \left(\bar{\Sigma}_1^2 + \alpha I\right)^{-2} \left(\alpha^2 I - \eta^2 \bar{\Sigma}_1^2\right) \bar{b}_1.$$

We can thus rewrite the secular equation $g(\alpha)$ in terms of α and β as

$$g(\alpha, \beta) = \bar{g}(\alpha) + \beta \frac{\alpha^2 - \eta^2\sigma_n^2}{(\sigma_n^2 + \alpha)^2}.$$

We note that when $\beta = 0$, we have $g(\alpha, 0) = \bar{g}(\alpha)$. Also note that $\bar{g}'(\alpha, 0) > 0$ when α lies in the interval $(\max(-\sigma_{n-1}^2, -\eta^2), \infty)$. Let $\alpha_1(\beta)$ denote the unique root in the interval $(-\sigma_n^2, \eta\sigma_1]$ and $\alpha_2(\beta)$ denote the rightmost root in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ of $g(\alpha, \beta)$. Also let $y_1(\beta)$ denote the stationary point $V^T\hat{x}$ corresponding to $\alpha_1(\beta)$ and similarly for $y_2(\beta)$ corresponding to $\alpha_2(\beta)$.

When $\bar{g}(-\sigma_n^2) < 0$, we note that neither $\alpha_1(\beta)$ nor $\alpha_2(\beta)$ converge to $-\sigma_n^2$ as β goes to zero. As already observed, at $\beta = 0$ we have that $g'(\alpha, 0) > 0$ when α lies in the interval $(\max(-\sigma_{n-1}^2, -\eta^2), \infty)$, and since $\bar{g}(-\sigma_n^2) < 0$ this implies that $g'(\alpha, 0) > 0$ when α lies in the interval $(\max(-\sigma_{n-1}^2, -\eta^2), -\sigma_n^2)$. Thus we know that $y_2(\beta)$ does not exist at $\beta = 0$ and thus it must not exist for some open neighborhood around $\beta = 0$. For $y_2(\beta)$ to be a candidate there must exist some value of β, say, β_2, for which $y_2(\beta)$ first exists. At the point β_2, $\alpha_2(\beta_2)$ must be at least a double root, and thus the slope of $g(\alpha_2(\beta))$ must be zero at β_2. From section 12, we note that $\alpha_2(\beta_2)$ cannot be the α^o, so we note that we must have $\|y_2(\beta_2)\|^2 \geq \|y_1(\beta_2)\|^2$.

We now proceed with the case when $\bar{g}(-\sigma_n^2) \geq 0$, and we will then show that in both cases $\|y_2(\beta)\|^2$ gets larger as β increases, while $\|y_1(\beta)\|^2$ decreases. It is easy to note from the form of $g(\alpha)$ that

$$\lim_{\beta \to 0+} \alpha_1(\beta) = -\sigma_n^2 = \lim_{\beta \to 0+} \alpha_2(\beta)$$

when $\bar{g}(-\sigma_n^2) \geq 0$. We now proceed to show that

$$\lim_{\beta \to 0+} |y_{1,i}(\beta)| = |y_{1,i}(0)| = |y_{2,i}(0)| = \lim_{\beta \to 0+} |y_{2,i}(\beta)|, \quad 1 \leq i \leq n.$$

First observe that this is trivially true for $i \neq n$. Next we note that $\bar{g}(\alpha)$ is continuous at $\alpha = -\sigma_n^2$; thus

$$\begin{aligned}
\lim_{\beta \to 0+} \left(y_{2,n}(\beta)^2 - y_{1,n}(\beta)^2\right) = \lim_{\beta \to 0+} \Bigg(& \frac{\sigma_n^2}{\alpha_2(\beta)^2 - \eta^2\sigma_n^2} \frac{\alpha_2(\beta)^2 - \eta^2\sigma_n^2}{\left(\alpha(\beta) + \sigma_n^2\right)^2} \beta \\
& - \frac{\sigma_n^2}{\alpha_1(\beta)^2 - \eta^2\sigma_n^2} \frac{\alpha_1(\beta)^2 - \eta^2\sigma_n^2}{\left(\alpha(\beta) + \sigma_n^2\right)^2} \beta \Bigg) \\
= \frac{1}{\sigma_n^2 - \eta^2} & \\
& \lim_{\beta \to 0+} \Bigg(\frac{\alpha_2(\beta)^2 - \eta^2\sigma_n^2}{\left(\alpha_2(\beta) + \sigma_n^2\right)^2} \beta + \bar{g}(\alpha_2(\beta), \beta) \\
& \qquad - \frac{\alpha_1(\beta)^2 - \eta^2\sigma_n^2}{\left(\alpha_1(\beta) + \sigma_n^2\right)^2} \beta - \bar{g}(\alpha_1(\beta), \beta) \Bigg)
\end{aligned}$$

$$= \frac{1}{\sigma_n^2 - \eta^2} \lim_{\beta \to 0+} \left(g(\alpha_2(\beta), \beta) - g(\alpha_1(\beta), \beta) \right)$$
$$= 0.$$

We note that we have shown what we desired to and therefore, $\|y_1(\beta)\|$ and $\|y_2(\beta)\|$ are continuous for $\beta \geq 0$, with $\|y_1(0)\| = \|y_2(0)\|$.

We now examine the derivative of the cost function, $\|x\|^2$, with respect to β. We will use this to show that in both cases $\|y_1(\beta)\|$ is less than $\|y_2(\beta)\|$ for all $\beta \geq 0$. The derivative is

$$\frac{d \left\| x(\alpha(\beta)) \right\|^2}{d\beta} = \frac{\sigma_n^2}{\left(\sigma_n^2 + \alpha(\beta) \right)^2} - 2 \frac{d\alpha(\beta)}{d\beta} b_1^T \left(\Sigma_1^2 + \alpha(\beta) I \right)^{-3} \Sigma_1^2 b_1.$$

We need to calculate the derivative of $\alpha(\beta)$ with respect to β, so we take the derivative of $g(\alpha(\beta)) = 0$:

$$0 = \frac{dg(\alpha(\beta))}{d\beta}$$
$$= \frac{\alpha(\beta)^2 - \eta^2 \sigma_n^2}{\left(\alpha(\beta) + \sigma_n^2 \right)^2} + 2 \left(\alpha(\beta) + \eta^2 \right) \frac{d\alpha(\beta)}{d\beta} \left(b_1^T \left(\Sigma_1^2 + \alpha(\beta) I \right)^{-3} \Sigma_1^2 b_1 \right).$$

Solving for the derivative of $\alpha(\beta)$ with respect to β yields

$$\frac{d\alpha(\beta)}{d\beta} = - \frac{\alpha(\beta)^2 - \eta^2 \sigma_n^2}{2 \left(\alpha(\beta) + \eta^2 \right) \left(\sigma_n^2 + \alpha(\beta) \right)^2 \left(b_1^T \left(\Sigma_1^2 + \alpha(\beta) I \right)^{-3} \Sigma_1^2 b_1 \right)}.$$

Substituting this into the derivative of $\|x\|^2$ with respect to β we obtain

$$\frac{d \left\| x(\alpha(\beta)) \right\|^2}{d\beta} = \frac{\sigma_n^2}{\left(\sigma_n^2 + \alpha(\beta) \right)^2} + \frac{\alpha(\beta)^2 - \eta^2 \sigma_n^2}{\left(\alpha(\beta) + \eta^2 \right) \left(\sigma_n^2 + \alpha(\beta) \right)^2}.$$

Simplifying this we get

$$\frac{d \left\| x(\alpha(\beta)) \right\|^2}{d\beta} = \frac{\alpha(\beta)}{\left(\alpha(\beta) + \eta^2 \right) \left(\alpha(\beta) + \sigma_n^2 \right)}.$$

Clearly, for increasing β we have that $\alpha_1(\beta)$ decreases the cost function when $\alpha_1(\beta) < 0$, while $\alpha_2(\beta)$ increases the cost function for all β. When $0 \leq \alpha_1(\beta) \leq \eta\sigma_n$ we have that $d\alpha(\beta)/d\beta \geq 0$ and we note that the cost is increasing for both $y_1(\beta)$ and $y_2(\beta)$. Since the cost is increasing for $y_1(\beta)$ when $0 \leq \alpha_1(\beta) \leq \eta\sigma_n$, we know that $\|y_1(\beta)\|^2 \leq \|y_1(\eta\sigma_n)\|^2$ on this interval. Additionally, note that for $\alpha_1(\beta)$ in the interval $[\eta\sigma_n, \eta\sigma_1]$ we have $d\alpha(\beta)/d\beta \leq 0$ and the cost increases with increasing β. Note that while these observations are true for $[\eta\sigma_n, \infty]$, we specify the interval $[\eta\sigma_n, \eta\sigma_1]$ because the root cannot lie in $[\eta\sigma_1, \infty]$. Observe that we now have $\|y_1(\beta)\|^2 \leq \|y_1(\eta\sigma_n)\|^2$ when $\alpha_1(\beta)$ is in the interval $[\eta\sigma_n, \eta\sigma_1]$. Thus the maximum value of the cost, when $\alpha_1(\beta)$ is in the interval $[\eta\sigma_n, \eta\sigma_1]$, occurs at $\beta = \eta\sigma_n$. We can easily find the maximum rate of change for the cost, when $\alpha_1(\beta)$ is in the interval $[0, \eta\sigma_1]$, to be

$$\max \frac{d \left\| x(\alpha(\beta)) \right\|^2}{d\beta} = \frac{\eta\sigma_n}{\left(\eta\sigma_n + \eta^2 \right) \left(\eta\sigma_n + \sigma_n^2 \right)}.$$

Simplifying we obtain

$$\max \frac{d\left\|x(\alpha(\beta))\right\|^2}{d\beta} = \frac{1}{(\eta+\sigma_n)^2}.$$

We can do similar calculation for the interval $(\max(-\eta^2, -\sigma_{n-1}^2), -\sigma_{n-1}^2)$ and we find that the minimum increase in the cost occurs at $\beta = -\eta\sigma_n$ and is given by

$$\min \frac{d\left\|x(\alpha(\beta))\right\|^2}{d\beta} = \frac{1}{(\eta-\sigma_n)^2}.$$

We now note that the maximum rate of increase for $y_1(\beta)$ is less than the minimum rate of increase for $y_2(\beta)$, and for β sufficiently small, we have $\|y_1(\beta)\| \leq \|y_2(\beta)\|$. We can now easily see that $\|y_1(\beta)\| \leq \|y_2(\beta)\|$ for all β; thus α_2 cannot be the global minimum.

We now consider the third candidate zero, namely, $-\sigma_{n-1}^2$. We note that for it to be a candidate we must have that $b_{1,n-1} = 0$ and $g(-\sigma_{n-1}^2) \geq 0$. We observe that similar to what we saw in Appendix B, the minimum on the interval $(-\sigma_{n-2}^2, -\sigma_n^2)$ must occur between the second to the rightmost and the rightmost roots of the secular equation on the interval. Recall in that section the only options were the roots themselves, but in this case there is also the possibility of $-\sigma_{n-1}^2$. Note that if $-\sigma_{n-1}^2$ is not one of the two rightmost roots on the interval $(-\sigma_{n-2}^2, -\sigma_n^2)$, then it cannot be the global minimum. We already know the rightmost root, designated α_2, is not the global minimum, and additionally the second most right root cannot be the global minimum since the slope of $g(\alpha)$, is not negative at this point.

We now reintroduce the parameter $\gamma = \|b_{1,n-1}\|^2$ and we will consider a continuity argument on γ similar to the continuity argument presented in this section. Since the argument is very similar to the one we constructed, we will again only sketch the details here. Note that for $\gamma \neq 0$ we have multiple roots in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$, none of which are the global minimum. As γ goes to zero we make all of the roots move to the left, and all but the rightmost either reaches $-\sigma_{n-1}^2$ or pops off the real line as $\gamma \to 0$, since $g(-\sigma_{n-1}^2) \geq 0$. The derivative of the cost with respect to γ can be seen to be negative in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ by the following method. First take the derivative and note there appears the term $d\alpha(\gamma)/d\gamma$, which we solve for by taking the derivative of $g(\alpha(\gamma)) = 0$ with respect to γ. Substituting back in and simplifying we see that as γ increases the cost decreases in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ and thus the x which corresponds to the root which appears in the interval when $\gamma \neq 0$ has a lower cost than the x which corresponds to $-\sigma_{n-1}^2$. That root is not a global minimum, however, and so neither can the root be at $-\sigma_{n-1}^2$. We can thus exclude the possibility that $-\sigma_{n-1}^2$ is α^o, and we are done.

19. Case 6: $\eta > \sigma_n$, $\|b_{1,n-k+1}\| \neq 0$, $\sigma_n = \sigma_{n-k+1}$. We again claim that there is a unique root, α_1, in the interval $(-\sigma_n^2, \eta\sigma_1]$ and it is the global minimum.

The existence of a root , α_1, in the interval $(-\sigma_n^2, \eta\sigma_1]$ follows from the observation that

$$\lim_{\alpha \to -\sigma_n^2 +} g(\alpha) = -\infty,$$
$$\lim_{\alpha \to \eta\sigma_1} g(\alpha) \geq 0.$$

Uniqueness is established by the same method as in section 14.

TABLE 3
Degeneracy conditions.

$\eta < \sigma_n$ and $b^T(I - A(A^TA - \eta^2 I)^{-1}A^T)b \leq 0$
$\eta = \sigma_n$, $b_{1,n} = 0$, and $\bar{b}_1^T(I - \bar{\Sigma}_1^2(\bar{\Sigma}_1 - \eta^2 I)^{-1})\bar{b}_1 \leq 0$
$\eta = \sigma_n$, $b_{1,n} \neq 0$
$\eta > \sigma_n$

Since $\|b_{1,n-k+1}\| \neq 0$, we cannot have $\alpha = \alpha_3 = -\sigma_n^2$. Note that the second order condition gives us the additional requirement that $\alpha \geq -\sigma_n^2$. Since $\alpha \geq -\sigma_n^2$ then trivially we do not have additional roots to worry about. The only candidate is thus the unique root, α_1, in the interval $(-\sigma_n^2, \eta\sigma_1]$.

20. Summary of results. The problem we have been considering is

$$\min_{x \in \mathcal{R}^n} \min_{\|E\| \leq \eta} \|(A + E)x - b\|,$$

where A is an $m \times n$ real matrix and b is an n-dimensional real column vector. We assume that the problem is degenerate and in particular that there exists an x such that $\eta\|x\| \geq \|Ax - b\|$. Degeneracy can be easily checked as outlined in Table 3. To obtain a solution to the degenerate problem we consider the optimization problem

$$\min_{\|Ax-b\| \leq \eta\|x\|} \|x\|.$$

The SVD of A is given by

$$A = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} V^T,$$

and we define $b_1 = U_1^T b$ and $b_2 = U_2^T b$. When $b_{1,n} = 0$ if σ_n is unique or $\|b_{1,n-k+1,n}\| = 0$ if σ_n is of multiplicity k, we can partition Σ_1 as

$$\Sigma_1 = \begin{pmatrix} \bar{\Sigma}_1 & 0 \\ 0 & \sigma_n I \end{pmatrix}.$$

We similarly partition b_1 into $\bar{b}_1$ and $b_{1,n} = 0$. The secular equation is given by

$$g(\alpha) = b_2^T b_2 + b_1^T(\Sigma_1^2 + \alpha I)^{-2}(\alpha^2 I - \eta^2 \Sigma_1^2)b_1.$$

Given these definitions, the solution to the problem is given in Table 4. Note that to find the unique root of the secular equation, $g(\alpha)$, in the interval specified can be easily and quickly done by a method such as bisection or Newton's method.

TABLE 4
Solution to the problem.

Condition	Solution
$\eta > \sigma_n$, $\sigma_n < \sigma_{n-1}$, $b_{1,n} = 0$, $g(-\sigma_n^2) \geq 0$	$x = V \begin{bmatrix} \left(\bar{\Sigma}_1^2 - \sigma_n^2 I\right)^{-1} \bar{\Sigma}_1 \bar{b}_1 \\ \pm\sqrt{\frac{g(-\sigma_n^2)}{\eta^2 - \sigma_n^2}} \end{bmatrix}$
$\eta > \sigma_n$, $\sigma_n = \sigma_{n-k+1}$, $\|b_{1,(n-k+1,n)}\| = 0$, $g(-\sigma_n^2) \geq 0$	$\hat{x} = V \begin{bmatrix} \left(\bar{\Sigma}_1^2 - \sigma_n^2 I\right)^{-1} \bar{\Sigma}_1 \bar{b}_1 \\ r\Theta \end{bmatrix}$ $r = \sqrt{\frac{g(-\sigma_n^2)}{\eta^2 - \sigma_n^2}}$ $\|\Theta\| = 1$
else	$x = (A^T A + \alpha I)^\dagger A^T b$ $\alpha_1 \in [\max(-\sigma_n^2, -\eta^2), \eta\sigma_1]$ such that $g(\alpha_1) = 0$

21. Restricted perturbations. We have so far considered the case in which all the columns of the A matrix are subject to perturbations. It may happen in practice, however, that only selected columns are uncertain, while the remaining columns are known precisely. This situation can be handled by the approach of this paper as we now clarify.

Given $A \in \Re^{m \times n}$, we partition it into block columns,

$$A = [\, A_1 \quad A_2 \,],$$

and assume, without loss of generality, that only the columns of A_2 are subject to perturbations while the columns of A_1 are known exactly. We then pose the following problem:

Given $A \in \Re^{m \times n}$, with $m \geq n$ and A full rank, $b \in \Re^m$, and nonnegative real number η_2, determine $\hat{x}$ such that

$$\min_{\hat{x}} \min_{\|\delta A_2\| \leq \eta_2} \{\|[\, A_1 \quad A_2 + \delta A_2 \,]\, \hat{x} - b\|\}. \tag{21.1}$$

If we partition $\hat{x}$ accordingly with A_1 and A_2, say,

$$\hat{x} = \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix},$$

then we can write

$$\|[\, A_1 \quad A_2 + \delta A_2 \,]\, \hat{x} - b\| = \|A\hat{x} - b + \delta A_2 \hat{x}_2\|.$$

Assuming the fundamental condition for this case, which is

$$\eta_2 \|\hat{x}_2\| \geq \|Ax - b\|,$$

and following the development of section 6 we conclude the problem is equivalent to

$$\min_{\|Ax-b\|^2 = \eta_2^2 \|x_2\|^2} \|x\|^2.$$

We note that we can rewrite the constraint as

$$\|Ax - b\|^2 + \eta_2^2\|x_1\|^2 = \eta_2^2\|x_2\|^2 + \eta_2^2\|x_1\|^2,$$

which becomes

$$\left\| \begin{bmatrix} A_1 & A_2 \\ \eta_2 I & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|^2 = \eta_2^2\|x\|^2.$$

We now define the following:

$$\tilde{A} = \begin{bmatrix} A_1 & A_2 \\ \eta_2 I & 0 \end{bmatrix}$$

and

$$\tilde{b} = \begin{bmatrix} b \\ 0 \end{bmatrix}.$$

The problem thus becomes

$$\min_{\|\tilde{A}x - \tilde{b}\|^2 = \eta_2^2\|x\|^2} \|x\|^2,$$

which is easily seen to be of the same form as our original problem, though of slightly larger dimension. This can thus be solved by the method discussed earlier in this paper.

Appendix A. Piecewise convexity of $\|x(\alpha)\|$.

We now show that $\|x(\alpha)\|^2$ is strictly convex in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$, which will allow us to show that only the zero closest to $-\sigma_n^2$ can correspond to a potential candidate for the global minimum.

We have that

$$\|x(\alpha)\|^2 = b_1^T \Sigma_1^2 (\Sigma_1^2 + \alpha I)^{-2} b_1.$$

Differentiating once with respect to α we get

$$\frac{d}{d\alpha}\|x(\alpha)\|^2 = -2b_1^T \Sigma_1^2 (\Sigma_1^2 + \alpha I)^{-3} b_1.$$

Differentiating once more we get

$$\frac{d^2}{d\alpha^2}\|x(\alpha)\|^2 = 6b_1^T \Sigma_1^2 (\Sigma_1^2 + \alpha I)^{-4} b_1,$$

from which we can conclude that $\|x(\alpha)\|^2$ is strictly convex on the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ and hence that it has a unique minimum on that interval.

Appendix B. Rightmost root.

We now show that of all the roots in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$ only the rightmost one can possibly correspond to the global minimum.

Let $\alpha_0, \ldots, \alpha_l$ denote the zeros of the secular equation $g(\alpha)$ in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$, in increasing order; that is,

$$-\sigma_{n-1}^2 < \alpha_0 < \alpha_1 < \cdots < \alpha_l < -\sigma_n^2.$$

From the result in section 12 we know that only the roots corresponding to negative slopes of the secular equation can correspond to local minima. Since

$$\lim_{\alpha \to -\sigma_n^2 -} g(\alpha) = -\infty,$$

it follows that

$$g'(\alpha_l) < 0 \quad \text{and} \quad g'(\alpha_{l-1}) > 0.$$

(We ignore the degenerate multiple root cases for now as the argument can be extended to them by continuity.)

Now there are two possibilities. Either $\|x(\alpha_l)\| \leq \|x(\alpha_{l-1})\|$ or not. The first case implies that $\|x(\alpha_{i+1})\| < \|x(\alpha_i)\|$ due to the convexity of $\|x(\alpha)\|$ on $(-\sigma_{n-1}^2, -\sigma_n^2)$.

For the second case we have that $\|x(\alpha_{l-1})\| < \|x(\alpha_l)\|$. We need to show this implies $\|x(\alpha_{l-1})\| < \|x(\alpha)\|$ for $-\sigma_{n-1}^2 < \alpha < -\alpha_{l-1}$, and that this is not the global minimum. Toward this end we take the derivative of $x(\alpha)$ with respect to α and get

$$\frac{dx(\alpha)}{d\alpha} = -\left(A^T A + \alpha I\right)^{-1} x(\alpha).$$

We have already shown that $\|x(\alpha)\|$ is convex on this interval, and thus it suffices to find if the derivative of $\|x(\alpha)\|^2$ with respect to α is negative at α_{l-1}, which shows that $x(\alpha)$ is then decreasing. We note that the derivative of $\|x(\alpha)\|^2$ is obtained by premultiplying the derivative of $x(\alpha)$ by $x(\alpha)^T$. To do the analysis we use the SVD of A and thus have

$$\frac{d\|x(\alpha)\|^2}{d\alpha} = -b_1^T \Sigma_1^2 \left(\Sigma_1^2 + \alpha I\right)^{-3} b_1.$$

We note that the matrix in parenthesis is indefinite and thus we must determine if the expression is negative or not at $\alpha = \alpha_{l-1}$. To do this we consider another function whose derivative we have already examined. Consider the constraint function, $\|Ax - b\|^2 - \eta^2 \|x\|^2$, and since at $\alpha = \alpha_{l-1}$ we are entering the infeasible region for increasing α, the derivative of the constraint must be positive. This condition can be expressed as

$$2(\alpha_{l-1} + \eta^2) b_1^T \Sigma_1^2 \left(\Sigma_1^2 + \alpha I\right)^{-3} b_1 > 0.$$

We note that $2(\alpha_{l-1} + \eta^2) > 0$, thus the condition is

$$b_1^T \Sigma_1^2 \left(\Sigma_1^2 + \alpha I\right)^{-3} b_1 > 0.$$

This trivially gives us

$$\frac{d\|x(\alpha)\|^2}{d\alpha} < 0,$$

and thus $x(\alpha)$ must be decreasing at $\alpha = \alpha_{l-1}$ for increasing α. Applying convexity to this result gives us $\|x(\alpha_{l-1})\| < \|x(\alpha)\|$ for $-\sigma_{n-1}^2 < \alpha < -\alpha_{l-1}$, and thus the minimum feasible value for $x(\alpha)$ on $-\sigma_{n-1}^2 < \alpha < -\sigma_n^2$ is $x(\alpha_{l-1})$.

Now since $x(\alpha_{l-1})$ does not correspond to a local minimum it follows that there is a neighborhood of $x(\alpha_{l-1})$ of the constraint surface such that in this neighborhood we have $\|x\| < \|x(\alpha_{l-1})\|$. Thus since $x(\alpha_{l-1})$ does not correspond to a local minimum, we can discard it from further consideration, since it is not the global minimum. Either way we are down to only the rightmost in the interval $(-\sigma_{n-1}^2, -\sigma_n^2)$.

REFERENCES

[1] S. Chandrasekaran, G. H. Golub, M. Gu, and A. H. Sayed, *Parameter estimation in the presence of bounded data uncertainties*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 235–252.

[2] S. Chandrasekaran, G. H. Golub, M. Gu and A. H. Sayed, An efficient algorithm for a bounded errors-in-variables model, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 839–859.

[3] G. H. Golub and C. F. Van Loan, *An analysis of the total least squares problem*, SIAM J. Numer. Anal., 17 (1980), pp. 883–893.

[4] S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*, Frontiers Appl. Math. 9, SIAM, Philadelphia, 1991.

[5] B. Walden, R. Karlson, and J. Sun, *Optimal backward perturbation bounds for the linear least squares problem*, Numer. Linear Algebra Appl., 2 (1995), pp. 271–286.