Backward Error Estimation

S. Chandrasekaran[*] E. Gomez [†] Y. Karant [‡] K. E. Schubert[§]

Abstract

Estimation of unknowns in the presence of noise and uncertainty is an active area of study, because no method handles all cases well, or even satisfactorily. The basic problem considered is the linear system, $Ax \approx b$, where A and b are given matrices with noise and uncertainty from measurements or modeling. The goal is to get the "best" estimate of x. The problem is useful in a wide variety of situations since it covers how to invert a matrix that contains uncertainty.

Mainstream methods like least squares and total least squares fail dramatically when A is ill-conditioned. Other methods like Tikhonov regression, ridge regression, and min max (bounded data uncertainty) provide robustness at the cost of fine details (by reducing $\|x\|$). This paper presents a new family of regression methods based off the backward error criteria which can add robustness and when possible enhances fine details (by increasing $\|x\|$).

This paper covers the motivation, proof, and application of backward error estimation. The resulting algorithms are compared against existing methods in numerical examples from image processing.

This paper is concerned with the estimation of unknowns that are related to some measurements by a linear model that is subject to uncertainty. Consider the set of linear equations, $Ax = b$, where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ are given. The goal is to calculate the value of $x \in \mathbb{R}^n$. If the equation is exact and A is not singular, the solution can be readily found by a variety of techniques, such as taking the QR factorization of A.

$$\begin{aligned} Ax &= b \\ QRx &= b \\ Rx &= Q^T b \end{aligned}$$

The last equation can be solved for x by back-substitution, since R is upper triangular. Given errors in modeling, estimation, and numeric representation the equality rarely holds. The least squares technique directly uses techniques like the QR factorization, by considering all the errors to be present in b. A more realistic appraisal of the system, considers errors in both A and b. Numerous methods exist for describing the errors in A and b, such as

1. bounding the norms of the errors in A and b,
2. constraining the errors in A and b to some structure,
3. partitioning A, b, and their corresponding errors, then placing bounds on the norms of each partition of the errors.

Combinations of the methods to describe the errors are also considered by some techniques. The description of the errors and constraints is one of the two fundamental ways a technique is specified for the linear model $Ax \approx b$. The other fundamental way of describing a method is to specify the cost function used to select the best value for x.

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

[†] Computer Science Department, California State University, San Bernardino, CA 92407-2397 (egomez@csci.csusb.edu).

[‡] Computer Science Department, California State University, San Bernardino, CA 92407-2397 (ykarant@csci.csusb.edu).

[§] Computer Science Department, California State University, San Bernardino, CA 92407-2397 (schubert@csci.csusb.edu).

Name	Cost Function	Constraint	Ψ
Least Squares	$\min_x \|Ax-b\|$		0
Total Least Squares	$\min_{[E_A \quad E_b]} \left\|[E_A \quad E_b]\right\|_F$	$(A+E_A)x = b+E_b$	$-\sigma_{n+1}^2 I$
Tikhonov	$\min_x \|Ax-b\| + \|Lx\|$		$L^T L$
Ridge Regression	$\min_x \left\| \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} x - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|$		λI
Min Max	$\min_x \max_{E_A \leq \eta, E_b \leq \eta_b} \|(A+E_A)x - (b+E_b)\|$		$\frac{\eta\|Ax-b\|}{\|x\|} I$
Min Min	$\min_x \min_{E_A \leq \eta, E_b \leq \eta_b} \|(A+E_A)x - (b+E_b)\|$		$-\frac{\eta\|Ax-b\|}{\|x\|} I$

Table 1: Techniques that can be expressed in the perturbed normal form of Eq. 1

Least squares considers the cost function, $\min_x \|Ax - b\|$. Total least squares minimizes the Frobenius norm of the errors in A and b (say E_A and E_b) subject to the resulting system being consistent, thus

$$\begin{aligned} &\min \left\|E_A \quad E_b\right\|_F \\ &\text{s.t.} \\ &\qquad (A+E_A)x = b + E_b. \end{aligned}$$

Other techniques consider directly minimizing the norm of $(A+E_A)x-(b+E_b)$ subject to some description of the errors as described above. For specialized situations other cost functions and error descriptions are considered, such as rational cost functions.

Despite numerous cost functions and error descriptions, the solution to many techniques can be expressed in a perturbed form of the normal equations, namely,

$$(A^T A + \Psi)x = A^T b, \tag{1}$$

with Ψ, frequently a scalar times the identity matrix. The exact form and value of Ψ is dependent on the technique and the data. Table 1 summarizes some of the techniques that can be expressed in the form of Eq. 1. For problems which can become degenerate (have multiple solutions) or can have different forms of Ψ, the most common form is listed in Table 1.

1 Motivation and Formulation

The backward error is used to prove that an algorithm has good numerical properties. The basic idea of the backward error analysis technique is to show the solution obtained is the exact solution of a nearby problem, by use of floating-point arithmetic error bounds. A method that does this is said to be a stable one. A method with a small backward error is not guaranteed accurate answers, but rather that the method used is a good one in the numerical sense. See [4] for a more complete treatment. It seems reasonable to ask then what the regression problem would be for a backward error criterion. The problem is thus stated

$$\min_x \max_{\|E\| \leq \eta} \frac{\|(A+E)x - b\|}{\|A\|\|x\| + \|b\|}.$$

It is important to note that this cost function is not convex, as most other cost functions are. This can be seen in Figure 1, which shows a region around the global minimum for a randomly generated matrix with 8 rows and 4 columns. Note in particular that a number of the minimums are close to the global minimum in cost (one within 2.5% of the cost). This problem is thus more difficult in nature than most, but it holds forth a potential of a numerically superior way of approaching the problem. To help understand the backward error problem better it will be considered as part of a family of four related cost functions. The problems are all rational cost functions and are generated by the presence or lack of the uncertainty in A (E) and the $\|b\|$ on the bottom. The costs functions are thus

1. $\min_x \max_{\|E\| \leq \eta} \frac{\|(A+E)x-b\|}{\|A\|\|x\|+\|b\|}$,

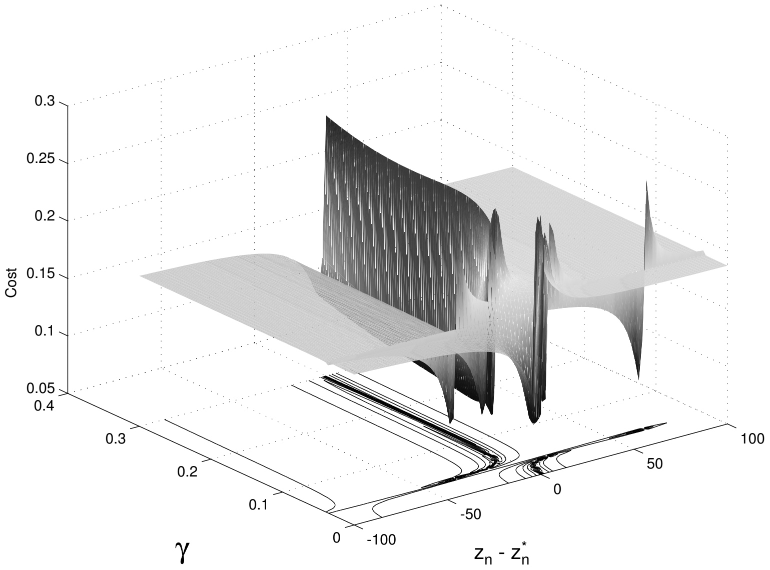


Figure 1: Cost function, showing the many singularities and relative minima.

2. $\min_x \frac{\|Ax-b\|}{\|A\|\|x\|+\|b\|}$,

3. $\min_x \max_{\|E\|\leq\eta} \frac{\|(A+E)x-b\|}{\|A\|\|x\|}$,

4. $\min_x \frac{\|Ax-b\|}{\|A\|\|x\|}$.

The three additional problems are thus nominal models in some sense. First, consider how to handle the maximization of the uncertainty (for problems 1 and 3). Since E appears only in the numerator, the maximization is accomplished by maximizing the numerator. Similar to what is done in [1], the original problem is identical to

$$\min_x \frac{\|Ax-b\|+\eta\|x\|}{\|A\|\|x\|+\|b\|}$$

and the third problem is identical to

$$\min_x \frac{\|Ax-b\|+\eta\|x\|}{\|A\|\|x\|} = \min_x \left(\frac{\|Ax-b\|}{\|A\|\|x\|} + \frac{\eta}{\|A\|}\right).$$

Essentially, problem three and problem four have the same solution! The four problems are thus

1. $\min_x \frac{\|Ax-b\|+\eta\|x\|}{\|A\|\|x\|+\|b\|} = \min_x C_1(x)$,

2. $\min_x \frac{\|Ax-b\|}{\|A\|\|x\|+\|b\|} = \min_x C_2(x)$,

3. $\min_x \frac{\|Ax-b\|+\eta\|x\|}{\|A\|\|x\|} = \min_x C_3(x)$,

4. $\min_x \frac{\|Ax-b\|}{\|A\|\|x\|} = \min_x C_4(x)$.

The problems will be examined in reverse order (simple to complex).

2 Formulations Three and Four

Since the third and fourth cost functions will give the same answer, consider the simpler model (fourth cost function) to get both. The fourth cost function can be squared without changing the solution. The squared model only has a non-differentiable point at $x = 0$, and is expressed as

$$\bar{C}_4(x) = \frac{\|Ax - b\|^2}{\|A\|^2\|x\|^2}.$$

Begin by noting that for this function, $x = 0$ is not a possible answer since the cost for $x = 0$ is always more costly than say the least squares solution x_{LS}. The solution is thus always at a differentiable point. Assume that $b \notin \mathcal{R}(A)$, since if it is not, the solution is trivially identical to the least squares solution. For all $x \neq 0$, the gradient of $\bar{C}_4(x)$ with respect to x is

$$\nabla_x \bar{C}_4(x) \quad = \quad 2\frac{\left(A^TA - \frac{\|Ax-b\|^2}{\|x\|^2}I\right)x - A^Tb}{\|A\|^2\|x\|^2}.$$

The candidate solutions, x_{opt}, are then found by setting $\nabla_x \bar{C}_4(x) = 0$ and solving to find

$$x_{opt} = \left(A^TA - \frac{\|Ax_{opt} - b\|^2}{\|x_{opt}\|^2}I\right)^{-1} A^Tb.$$

Thus consider the parameterized family $x(\psi)$ given by

$$x(\psi) = \left(A^TA - \psi I\right)^{-1} A^Tb.$$

Define the regression parameter as $\psi = \frac{\|Ax-b\|^2}{\|x\|^2}$, and thus ψ_{opt} is given by

$$\psi_{opt} = \frac{\|Ax_{opt} - b\|^2}{\|x_{opt}\|^2} > 0. \tag{2}$$

An expression to calculate ψ_{opt} is needed. Rewrite the definition of ψ_{opt} and substitute the expression for x_{opt} found above.

$$\psi_{opt}\|\left(A^TA - \psi_{opt}I\right)^{-1} A^Tb\|^2 \quad = \quad \|A\left(A^TA - \psi_{opt}I\right)^{-1} A^Tb - b\|^2.$$

To simplify the expression above, introduce the singular value decomposition (SVD) of A as $A = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma & 0 \end{bmatrix}^T V^T$. Also introduce the notation $b_1 = U_1^Tb$ and $b_2 = U_2^Tb$. The expression becomes

$$\psi_{opt}b_1^T\left(\Sigma^2 - \psi_{opt}I\right)^{-1} b_1 - b_2^Tb_2 \quad = \quad 0.$$

Call the expression,

$$g_1(\psi) = \psi b_1^T\left(\Sigma^2 - \psi I\right)^{-1} b_1 - b_2^Tb_2, \tag{3}$$

the secular equation in keeping with the literature [1, 2]. Thus the value of ψ_{opt} is specified by the roots of the secular equation. To find the root several questions need to be answered. Does the root exist? Is the root unique? Is there an interval where the root occurs? Immediately note from the expression for ψ_{opt}, Equation 2, that it is greater than zero. The basic outline is to find an upper bound on the size of the ψ_{opt}, which is a local minimum. When the upper bound is found, establish uniqueness and finally existence.

Before proceeding further, note a simple relation that can be derived from $x(\psi) = \left(A^TA - \psi I\right)^{-1} A^Tb$ and will prove useful in our development. Note that this relation holds for all values of ψ.

$$x(\psi) = \left(A^TA - \psi I\right)^{-1} A^Tb \Rightarrow A^T(Ax(\psi) - b) \quad = \quad \psi x(\psi)$$

Now proceed with taking the second derivative of the cost,

$$\nabla_x^2 \bar{C}_4(x) \quad = \quad 2\frac{A^T A - \psi I - 2P_x\left(\psi - \frac{\|Ax-b\|^2}{\|x\|^2}\right)}{\|A\|^2\|x\|^2} - 2\frac{\nabla_x \bar{C}_4(x) x^T}{\|x\|^2}.$$

Note that P_x is the projection onto x and is given by $\frac{xx^T}{\|x\|^2}$. Since only the slope at the roots of the secular equation are of concern, that means $\nabla_x \bar{C}_4(x) = 0$ and $\psi = \frac{\|Ax-b\|^2}{\|x\|^2}$. The second derivative becomes $2(A^T A - \psi I)$, which is positive definite for $\psi < \sigma_n^2$, where σ_n is the smallest singular value of A. This means

$$\boxed{0 \le \psi \le \sigma_n^2.}$$

To show uniqueness, take the derivative of the secular equation,

$$g_1'(\psi) \quad = \quad b_1^T \Sigma^2 \left(\Sigma^2 - \psi I\right)^{-2} b_1 > 0.$$

The derivative is positive, so the root will be unique. Note that discontinuities exist, but not in the interval where the solution must lie. All that remains is then to show existence.

Begin by observing that $g_1(0) = -b_2^T b_2 \le 0$. For simplicity, assume that the smallest singular value of A is unique, the extension is obvious. Now if the n^{th} element of b_1, denoted $b_{1,n}$ is not zero then $\lim_{\psi \to \sigma_n^2} g_1(\psi) = \infty$. If $b_{1,n} \ne 0$ then trivially the root exists. If $b_{1,n} = 0$ then note that

$$g_1(\sigma_n^2) = \sigma_n^2 \bar{b_1}^T (\bar{\Sigma}^2 - \psi I)^{-1} \bar{b_1} - b_2^T b_2,$$

with

$$b_1 = \begin{bmatrix} \bar{b_1} \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \bar{\Sigma} & 0 \\ 0 & \sigma_n \end{bmatrix}.$$

If this number is non-negative, again the root exists. The question remaining is what happens when the number is negative? In this case look at the gradient of the cost function

$$\begin{bmatrix} \bar{\Sigma}^2 - \psi I & 0 \\ 0 & \sigma_n^2 - \psi \end{bmatrix} \begin{bmatrix} \bar{z} \\ z_n \end{bmatrix} = \begin{bmatrix} \bar{\Sigma}\bar{b_1} \\ 0 \end{bmatrix}$$

where,

$$\begin{bmatrix} \bar{x} \\ x_n \end{bmatrix} = v \begin{bmatrix} \bar{z} \\ z_n \end{bmatrix}.$$

Since $g_1(\sigma_n^2) < 0$ and $g_1'(\psi) > 0$, there is no root of $g_1(\psi)$ for $\psi < \sigma_n^2$. The global minimum root is less than or equal to σ_n^2, so $\psi = \sigma_n^2$. This means that $\bar{z} = (\bar{\Sigma}^2 - \sigma_n^2 I)^{-1}\bar{\Sigma}\bar{b_1}$. To find the value of z_n , substitute the values of ψ and $\bar{z}$ back into the cost function and find that

$$\bar{C}_4\left(v\begin{bmatrix} \bar{z} \\ z_n \end{bmatrix}\right) = \frac{\sigma_n^2}{\sigma_1^2} \frac{\sigma_n^4 \bar{b_1}^T \left(\bar{\Sigma}^2 - \psi I\right)^{-2} \bar{b_1} + b_2^T b_2 + \sigma_n^2 z_n^2}{\sigma_n^2 \bar{b_1}^T \bar{\Sigma}^2 \left(\bar{\Sigma}^2 - \psi I\right)^{-2} \bar{b_1} + \sigma_n^2 z_n^2}.$$

Only three possibilities exist, $z_n = 0$, $z_n = \pm\infty$, or z_n can be anything. For a rational function of the form

$$\frac{\alpha + \sigma_n^2 z_n^2}{\beta + \sigma_n^2 z_n^2},$$

the value of z_n is given by

$$z_n = \begin{cases} 0 & \beta - \alpha > 0 \\ \pm\infty & \beta - \alpha < 0 \\ * & \beta - \alpha = 0. \end{cases}$$

For this problem $\beta - \alpha = g_1(\sigma_n^2) < 0$ and thus $z_n = \pm\infty$. The entire solution is thus characterized.

3 Relation to TLS

Two regression problems hold predominance in estimation, least squares and total least squares. For formulations three and four, the regression parameter, ψ is greater than zero, which is the least squares regression parameter. The parameter ψ also has a relation to the total least squares regression parameter σ_{n+1}.

Start by noting an interesting bound on the size of the total least squares regression parameter. To see the bound, consider the TLS problem written as

$$\begin{bmatrix} A^T A & A^T b \\ b^T A & b^T b \end{bmatrix} \begin{bmatrix} x_{TLS} \\ -1 \end{bmatrix} = \sigma_{n+1}^2 \begin{bmatrix} x_{TLS} \\ -1 \end{bmatrix}.$$

The top line specifies the form of the solution as

$$x_{TLS} = \left(A^T A - \sigma_{n+1}^2 I\right)^{-1} A^T b.$$

The bottom line gives the secular equation for the TLS problem. This can be seen by inserting the form of solution into the bottom line.

$$\begin{aligned} b^T A x_{TLS} - b^T b &= -\sigma_{n+1}^2 \\ b^T A \left(A^T A - \sigma_{n+1}^2 I\right)^{-1} A^T b - b^T b &= -\sigma_{n+1}^2. \end{aligned}$$

Recall the SVD of A and the definitions of b_1 and b_2 that have been used in earlier sections,

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

This yields on rearranging

$$\sigma_{n+1}^2 \left(b_1^T (\Sigma^2 - \sigma_{n+1}^2 I)^{-1} b_1 + 1\right) = b_2^T b_2. \tag{4}$$

Recalling that $\sigma_{n+1} \leq \sigma_n$ it can be seen that

$$b_1^T (\Sigma^2 - \sigma_{n+1}^2 I)^{-1} b_1 \geq 0,$$

and thus

$$b_1^T (\Sigma^2 - \sigma_{n+1}^2 I)^{-1} b_1 + 1 \geq 1.$$

Finally arrive at the desired upper bound of the TLS parameter,

$$\boxed{\sigma_{n+1}^2 \leq b_2^T b_2.}$$

Now proceed to see how the regression parameter from Section 2 compares with the TLS parameter, σ_{n+1}. To do so we subtract Equation 4 from Equation 3 and obtain

$$b_1^T (\Sigma^2 - \psi I)^{-1} (\Sigma^2 - \sigma_{n+1}^2 I)^{-1} \left(\psi(\Sigma^2 - \sigma_{n+1}^2 I) - \sigma_{n+1}^2 (\Sigma^2 - \psi I)\right) b_1 = \sigma_{n+1}^2.$$

Recall that $0 \leq \sigma_{n+1} \leq \sigma_n$ and $0 \leq \psi \leq \sigma_n$ so

$$\begin{aligned} \Sigma^2 - \psi I &\geq 0 \\ \Sigma^2 - \sigma_{n+1}^2 I &\geq 0. \end{aligned}$$

Thus the following results

$$\begin{aligned} \psi(\Sigma^2 - \sigma_{n+1}^2 I) - \sigma_{n+1}^2 (\Sigma^2 - \psi I) &\geq 0 \\ \psi - \sigma_{n+1}^2 &\geq 0. \end{aligned}$$

The final result is

$$\psi \geq \sigma_{n+1}^2.$$

Formulations three and four always deregularize more than TLS. This gives the even better bounds for ψ of

$$\boxed{\sigma_{n+1}^2 \leq \psi \leq \sigma_n^2.}$$

4 Formulation Two

The main difficulty in solving the general problem (problem 1) is the denominator. In particular the addition of the $\|b\|$ in the denominator adds considerable complexity. It is reasonable to ask why bother with it, after all a solution exists without it. The main problem with formulations three and four are that they are always too optimistic. As x approaches zero from sufficiently close, the denominator goes to zero and thus the cost rises. It is never possible for this regression technique to return $x = 0$ even though there are times when that is the best choice from a physical standpoint. The constant addition of $\|b\|$ in the denominator of the cost prevents the cost function from going to infinity as x is near zero. Formulations one and two are thus more realistic than formulations three and four. It is important to note that for formulation two, the only time $x = 0$ is when $A^\dagger b = 0$, as this is the only way to have $\|Ax - b\| = \|A\|\|x\| + \|b\|$. In practical terms this will not happen so it will be assumed that $A^\dagger b \neq 0$ for this section and thus $x = 0$ is not a candidate. When $b = AA^\dagger b$, the choice of $x = A^\dagger b$ yields a cost of zero, so it is the solution. When $b \neq AA^\dagger b$, the only non-differentiable point has been ruled out so the solution is at a differentiable point.

The next step in solving the general problem (formulation) is to consider the next most difficult problem (formulation two). Formulation two is defined by the cost function

$$\min_x C_2(x) = \min_x \frac{\|Ax - b\|}{\|A\|\|x\| + \|b\|}.$$

As with the other problems, take the gradient with respect to x and by rearranging terms, find that

$$\nabla_x C_2(x) \quad = \quad \frac{A^T(Ax - b) - \psi_2 x}{\|Ax - b\|(\|A\|\|x\| + \|b\|)}$$

with $\psi_2 = \frac{\|Ax-b\|^2\|A\|}{\|x\|(\|A\|\|x\|+\|b\|)}$. By setting $\nabla_x C_2(x)$ equal to zero, this yields

$$x(\psi_2) \quad = \quad (A^T A - \psi_2 I)^{-1} A^T b.$$

The Hessian is given by

$$\nabla_x^2 C_2(x) \quad = \quad \frac{1}{\|Ax - b\|(\|A\|\|x\| + \|b\|)} \left(A^T A - \psi_2 I + \psi_2 \frac{\|b\|xx^T}{(\|A\|\|x\| + \|b\|)\|x\|^2} \right).$$

Denote the singular value decomposition of A as before, with the smallest singular value of A given by σ_n. First note that when $\psi_2 \leq \sigma_n^2$ then the Hessian is positive semidefinite.

The Hessian will be positive semidefinite if $A^T A - \psi_2 I + \psi_2 \frac{\|b\|xx^T}{(\|A\|\|x\|+\|b\|)\|x\|^2}$ is positive semidefinite. Using the SVD and denoting $z = v^T x$ the condition becomes $\Sigma^2 - \psi_2 I + \psi_2 \frac{\|b\|zz^T}{(\|A\|\|z\|+\|b\|)\|z\|^2}$ must be non-negative. By partitioning z into $z = [\bar{z}^T \quad z_n]^T$ and partitioning the remaining matrices similarly, the Hessian condition can be written as:

$$\begin{bmatrix} \bar{\Sigma}^2 - \psi I & 0 \\ 0 & \sigma_n^2 - \psi \end{bmatrix} + \psi_2 \frac{\|b\|}{(\|A\|\|z\| + \|b\|)\|z\|^2} \begin{bmatrix} \bar{z} \\ z_n \end{bmatrix} \begin{bmatrix} \bar{z} \\ z_n \end{bmatrix}^T$$

and the form of solution becomes

$$(\Sigma^2 - \psi_2 I)z \quad = \quad \Sigma^T b_1$$

with $b_1 = U_1 b$ as before. When $b_{1,n} = 0$ then either $z_n = 0$ or $\psi_2 = \sigma_n^2$. If $\psi_2 = \sigma_n^2$, trivially $\psi_2 \leq \sigma_n^2$. But note that if $z_n = 0$ then by the Hessian, $\psi_2 \leq \sigma_n^2$.

$$\boxed{b_{1,n} = 0 \quad \Rightarrow \quad \psi_2 \leq \sigma_n^2}$$

4.1 Perturbation Analysis

It has already been shown that when $b_{1,n} = 0$ that $\psi_2 \leq \sigma_n^2$, so it remains to be shown that this remains true when $b_{1,n} \neq 0$. Note from perturbation theory (for example Section 8.6.1 of [3]) that for the Hessian condition to be non-negative it must have either $\psi_2 \leq \sigma_{n-1}^2$ or $\psi_2 \leq \sigma_n^2 \frac{\|A\|\|z\|+\|b\|}{\|A\|\|z\|}$. In particular if the smallest singular value is a multiple singular value, so $\sigma_n = \sigma_{n-1}$, then trivially $\psi_2 \leq \sigma_n^2$. As a final observation, $\psi_2 = \sigma_n^2$ only when $b_{1,n} = 0$ so by continuity of ψ_2, in order to have $\psi_2 > \sigma_n^2$ it must first have $\psi_2 = \sigma_n^2$ when $b_{1,n} = 0$. It only remains to show that when $b_{1,n} \neq 0$ and σ_n is a unique singular value that $\psi_2 \leq \sigma_n^2$.

This is done by first performing a perturbation analysis on $b_{1,n}$. When $b_{1,n} = 0$, $\psi_2 = \sigma_n^2$, so when $b_{1,n} = \delta b_{1,n} \ll 1$, $\psi_2 = \sigma_n^2 \pm \delta\psi_2$. Note that by continuity there is always found a small value of $\delta b_{1,n}$ such that $\delta\psi_2 \ll \sigma_{n-1}^2 - \sigma_n^2$. To show that $\psi_2 < \sigma_n^2$ it needs to be shown that the cost for $\psi_2 = \sigma_n^2 + \delta\psi_2$ is greater than the cost for $\psi_2 = \sigma_n^2 - \delta\psi_2$. Proceed by examining the cost function and the first order condition of the cost function.

Note that for the partitioning

$$z(\psi_2) \quad = \quad \begin{bmatrix} \bar{z} \\ z_n \end{bmatrix}$$

the cost function can be rewritten as

$$C_2 = \frac{\left\| \begin{bmatrix} \bar{\Sigma} & 0 \\ 0 & \sigma_n \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \bar{z} \\ z_n \end{bmatrix} - \begin{bmatrix} \bar{b}_1 \\ b_{1,n} \\ b_2 \end{bmatrix} \right\|}{\|A\| \left\| \begin{bmatrix} \bar{z} \\ z_n \end{bmatrix} \right\| + \|b\|}.$$

The first order condition can likewise be written as

$$\begin{bmatrix} \bar{\Sigma}^2 - \psi_2 I & 0 \\ 0 & \sigma_n^2 - \psi_2 \end{bmatrix} \begin{bmatrix} \bar{z} \\ z_n \end{bmatrix} \quad = \quad \begin{bmatrix} \bar{\Sigma} & 0 \\ 0 & \sigma_n \end{bmatrix} \begin{bmatrix} \bar{b}_1 \\ b_{1,n} \end{bmatrix}.$$

Using the fact that $\psi_2 = \sigma_n^2 \pm \delta\psi_2$ it can be seen that $\mp\delta\psi_2 = \sigma_n^2 - \psi_2$. Define $\bar{D} = \bar{\Sigma} - \sigma_n^2 I$ then rewrite the first order condition as

$$\begin{bmatrix} \bar{D} \mp \delta\psi_2 I & 0 \\ 0 & \mp\delta\psi_2 \end{bmatrix} \begin{bmatrix} \bar{z} \\ z_n \end{bmatrix} \quad = \quad \begin{bmatrix} \bar{\Sigma} & 0 \\ 0 & \sigma_n \end{bmatrix} \begin{bmatrix} \bar{b}_1 \\ b_{1,n} \end{bmatrix}.$$

Since it was noted above that $\delta\psi_2 \ll \sigma_{n-1}^2 - \sigma_n^2$ and $\sigma_{n-1}^2 - \sigma_n^2$ is the smallest element of the diagonal matrix $\bar{D}$, approximate $\bar{D} \mp \delta\psi_2 I$ by $\bar{D}$. This results in

$$\begin{aligned} \begin{bmatrix} \bar{D} & 0 \\ 0 & \mp\delta\psi_2 \end{bmatrix} \begin{bmatrix} \bar{z} \\ z_n \end{bmatrix} \quad &= \quad \begin{bmatrix} \bar{\Sigma} & 0 \\ 0 & \sigma_n \end{bmatrix} \begin{bmatrix} \bar{b}_1 \\ b_{1,n} \end{bmatrix} \\ \begin{bmatrix} \bar{z} \\ z_n \end{bmatrix} \quad &= \quad \begin{bmatrix} \bar{D}^{-1}\bar{\Sigma}\bar{b}_1 \\ \frac{\sigma_n b_{1,n}}{\mp\delta\psi_2} \end{bmatrix}. \end{aligned}$$

From this note that the norm of z will not be affected by the the sign of $\mp\delta\psi_2$, thus only the numerator of the cost matters. Substituting the result into the cost function find

$$C_2 \quad = \quad \frac{\left\| \begin{bmatrix} \sigma_n^2(\bar{\Sigma} - \sigma_n^2 I)^{-1}\bar{b}_1 \\ \frac{\sigma_n^2 \pm \delta\psi_2}{\mp\delta\psi_2} b_{1,n} \\ -b_2 \end{bmatrix} \right\|}{\|A\|\|z\| + \|b\|}.$$

The only term where the sign of $\pm\delta\psi_2$ is significant is the second row of the numerator. When the sign is positive the second row is clearly larger than when the sign is negative. Thus the norm of the numerator will be larger when the sign is positive, so the cost is less when $\psi_2 < \sigma_n^2$ then when $\psi_2 > \sigma_n^2$. This eliminates the possibility that ψ_2 lies in the range $(\sigma_n^2, \sigma_{n-1}^2)$.

4.2 Final Case

There only remains the possible alternative of $\psi_2 = \sigma_{n-1}^2$, which will be disproven now. Consider the first derivative of the cost, with $z = \begin{bmatrix} \bar{z}^T & z_{n-1} & z_n \end{bmatrix}^T$. Partition the other matrices similarly to obtain

$$\begin{bmatrix} \bar{\Sigma}^2 - \psi_2 I & 0 & 0 \\ 0 & \sigma_{n-1}^2 - \psi_2 & 0 \\ 0 & 0 & \sigma_{n-1}^2 - \psi_2 \end{bmatrix} \begin{bmatrix} \bar{z} \\ z_{n-1} \\ z_n \end{bmatrix} = \begin{bmatrix} \bar{\Sigma}\bar{b} \\ \sigma_{n-1} b_{1,n-1} \\ \sigma_n b_{1,n} \end{bmatrix}.$$

For $\psi_2 = \sigma_{n-1}^2$ it is required that $b_{1,n-1} = 0$ so

$$\begin{bmatrix} \bar{\Sigma}^2 \sigma_{n-1}^2 I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \sigma_{n-1}^2 - \sigma_{n-1}^2 \end{bmatrix} \begin{bmatrix} \bar{z} \\ z_{n-1} \\ z_n \end{bmatrix} = \begin{bmatrix} \bar{\Sigma}\bar{b} \\ 0 \\ \sigma_n b_{1,n} \end{bmatrix}.$$

It has been shown that if $b_{1,n} = 0$ then $\psi_2 \leq \sigma_n^2$, so if $\psi_2 = \sigma_{n-1}^2$ then the determinant of the second derivative of the cost must be positive for all values of $b_{1,n} \neq 0$. Consider the second derivative of the cost partitioned as was done for the first derivative and with $\psi_2 = \sigma_{n-1}^2$.

$$\begin{bmatrix} \bar{\Sigma}^2 - \sigma_{n-1}^2 I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \sigma_n^2 - \sigma_{n-1}^2 \end{bmatrix} + \frac{\sigma_{n-1}^2 \|b\|}{(\|A\|\|z\| + \|b\|)\|z\|^2} \begin{bmatrix} \bar{z} \\ z_{n-1} \\ z_n \end{bmatrix} \begin{bmatrix} \bar{z} \\ z_{n-1} \\ z_n \end{bmatrix}^T$$

The determinant of this is always less than or equal to zero (proof of this is omitted due to length) in some neighborhood of $b_{1,n} = 0$, which means that it can never be the minimum at any point, since in order to be the minimum it must be the minimum at some point around $b_{1,n} = 0$. It is thus impossible for $\psi_2 = \sigma_{n-1}^2$.

4.3 Final Result

From what has been shown, the solution to problem 2 is given by

$$x(\psi_2) \quad = \quad (A^T A - \psi_2 I)^{-1} A^T b$$

where

$$\psi_2 \in [0, \sigma_n^2].$$

The value of ψ_2 can be found as the root of

$$\psi_2 - \frac{\|Ax(\psi_2) - b\|^2 \|A\|}{\|x(\psi_2)\|(\|A\|\|x(\psi_2)\| + \|b\|)}$$

in the given range. Since the Hessian is strictly positive in this range and for $\psi_2 = 0$ the equation is trivially less than zero, the root is unique in this range. The root can be found in n^2 time by using a root finding method like bisection or Newton's method.

5 Formulation One

Before examining the general solution, recall that a noted problem of the other formulations was that $x = 0$ was not a candidate or that the uncertainty could not make it happen. In the final formulation this problem is handled. The cost for $x = 0$ is 1, so for $x = 0$ to be the solution it must be true that for all $x \neq 0$ that

$$\begin{aligned} \|Ax - b\| + \eta\|x\| &\geq \|A\|\|x\| + \|b\| \\ \Rightarrow \quad \eta &\geq \sigma_1 + \frac{\|b\| - |\|Ax\| - \|b\||}{\|x\|}. \end{aligned}$$

If $\|b\| > \|Ax\|$ then rearranging yields $\eta \geq 2\sigma_1$. If $\|b\| \leq \|Ax\|$ then rearranging yields $\eta \geq 2\sigma_1$. Thus no matter what, for $x = 0$ to be the solution, $\eta \geq 2\sigma_1$. The point at which $x = 0$ is a candidate solution can be adjusted by changing the term, $\|A\|\|x\|$, in the denominator to $\alpha\|x\|$ for $0 < \alpha \leq \|A\|$. This does not alter the analysis but does permit smaller values of η to yield $x = 0$. Note that $\alpha = 0$ is excluded as this is the min max problem discussed in [1].

We now proceed to the form of solution and secular equation to find when the solution is not at a singular point. The cost function is

$$C_1(x) = \frac{\|Ax - b\| + \eta\|x\|}{\|A\|\|x\| + \|b\|}.$$

Taking the gradient and setting it equal to zero yields

$$x(\psi_1) = (A^TA + \psi_1 I)^{-1}A^Tb,$$

with

$$\begin{aligned}\psi_1 &= \frac{\|Ax-b\|}{\|x\|}\left(\eta - \frac{\|Ax-b\| + \eta\|x\|}{\|A\|\|x\| + \|b\|}\|A\|\right)\\ &= \frac{\|Ax-b\|}{\|x\|}\left(\eta - C_1(x)\|A\|\right).\end{aligned}$$

The Hessian of the cost when the gradient of the cost is zero (i.e. the Hessian of the cost for the candidate solutions) is given by

$$H_1 = A^TA + \left(\frac{\psi_1^2}{\|Ax-b\|^2} + \frac{\psi_1}{\|x\|^2}\right)xx^T + \psi_1 I.$$

For the solution to be a minimum, the Hessian must be positive. For the Hessian to be positive, $V_n^T H_1 V_n > 0$, where V_n is the right singular vector that corresponds to the smallest singular value, σ_n.

$$\begin{aligned}V_n^T H_1 V_n &= \sigma_n^2 + \left(\frac{\psi_1^2}{\|Ax-b\|^2} + \frac{\psi_1}{\|x\|^2}\right)\left(\frac{\sigma_n b_{1,n}}{\sigma_n^2 + \psi_1}\right)^2 + \psi_1\\ &= (\sigma_n^2 + \psi_1) + \frac{\psi_1^2 b_1^T(\Sigma^2 + \psi_1)^{-1}b_1\sigma_n^2 b_{1,n}^2}{\|Ax-b\|^2\|x\|^2(\sigma_n^2+\psi_1)^2} + \frac{\psi_1 b_2^T b_2 \sigma_n^2 b_{1,n}^2}{\|Ax-b\|^2\|x\|^2(\sigma_n^2+\psi_1)^2}\end{aligned}$$

Note that if $\psi_1 < -\sigma_n^2$ then all three terms are negative, thus the Hessian is not positive. The solution is thus in the interval

$$\boxed{\psi_1 \in [-\sigma_n^2, \infty).}$$

This is the solution, and it can be calculated by running down all the roots of $g(\psi_1) = \psi_1 - \frac{\|Ax-b\|}{\|x\|}\left(\eta - C_1(x)\|A\|\right)$ in the interval.

6 Numerical Examples

Blurring occurs often in images. For example atmospheric conditions, dust, or imperfections in the optics can cause a blurred image. Blurring is usually modelled as a Gaussian blur, which is a great smoothing filter. The Gaussian blur causes greater distortion on the corners, which is exactly where we do not want it. The component of a Gaussian blur with standard deviation, σ, in position, (i,j), is given by

$$G_{i,j} = e^{-\left(\frac{i-j}{\hat{\sigma}}\right)^2}.$$

If we go on the presumption that we do not know the exact blur that was applied (the standard deviation, $\hat{\sigma}$ unknown) we cannot expect to get the exact image back. While we realize that we will not be able to perfectly extract the original system, 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.

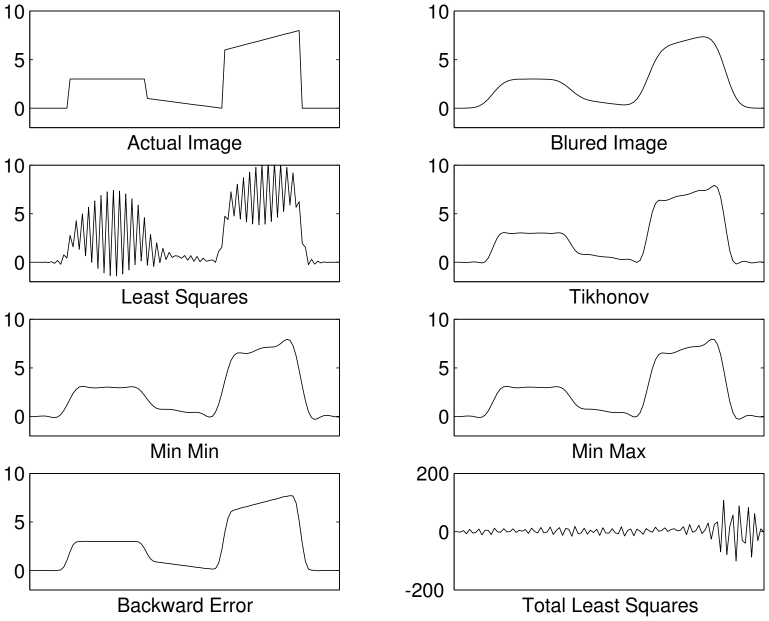


Figure 2: Skyline Problem

6.1 First Example

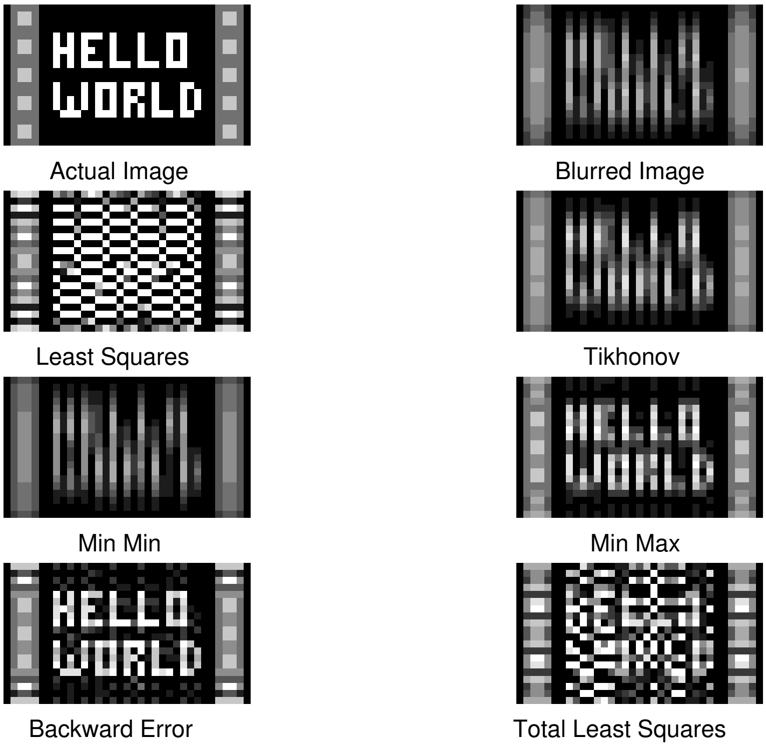
Consider a simple one dimensional “skyline” image that has been blurred. A “skyline” image is a one dimensional image that looks like a city skyline when graphed, and thus is the most basic image processing example. “Skyline” images involve sharp corners, and it is of key importance to accurately locate these corner transitions. The LS solution captures general trends, but still not acceptable, see Figure 2. The Tikhonov solution works well due to its increased robustness. Observe that the min min solution exhibits robustness as this is one of the cases where the problem is degenerate and it can either regularize or de-regularize. In this case the solution is regularized due to the relatively large uncertainty. The min max performs well due to its robustness. Most interestingly note that the backward error solution performs the best of all. It does an excellent job of finding the corners without sacrificing the edges. Finally, the TLS solution fails completely, yielding a result that is about two orders of magnitude off.

6.2 Second Example

The second example is a simple two-dimensional image processing application. A small picture with the grey-scale words, ‘HELLO WORLD’ of early programming fame, has been blurred. The image is 20x35 and the blur is done by a Gaussian blur matrix of size 20. The blur is not so strong that some of the features cannot be seen, and in particular one can see that there is writing but the specifics are hard to make out.

A key aspect of all of the regression techniques is selection of the regression parameter. In the suggested techniques, this is done semi-automatically. Semi-automatically because the error bound on the matrix must still be supplied, and it is not guaranteed to be known accurately. This becomes critical as the selection of the regression parameter is mostly influenced by this error bound. Select an error that is too large and data losses can result, select one too small and there will not be enough regulation or deregulation to improve the regression. The error bound was selected to be proportional to the 2-norm of the actual error.

Least squares solution is not readable at all, due to its lack of robustness. Tikhonov makes some small gains, but not enough to be useful. The min min solution makes no noticeable improvements. The min max technique generates a readable result. The BE solution is very readable, though obviously not perfect. Total least squares also makes the image worse, though you can almost make out the some rough letters so it is better than least squares in this case.



Actual Image

Blurred Image

Least Squares

Tikhonov

Min Min

Min Max

Backward Error

Total Least Squares

Figure 3: Hello World with $\eta = \|\delta A\|_2$

7 Conclusions

Several techniques have been proposed to handle uncertainty in estimation. Each has cases where it provides improvement, thus warranting their consideration. This paper proposes the backward error technique, which frequently outperforms other techniques when robustness is needed. For many runs of the numerical examples with different error bounds to simulate different assumptions made by the modeler, the BE technique ended up giving reasonable answers for a longer region. This suggests that the backward error technique is more robust than other techniques to errors. It must be noted that fine tuning of the perturbation error bound is very helpful, though, even for robust systems.

References

[1] S. Chandrasekaran, G. H. Golub, M. Gu, and A. H. Sayed, *Parameter Estimation in the Presence of Bounded Data Uncertainties*, SIMAX, 19 (1998), pp. 235–252.

[2] ——, *An Efficient Algorithm for a Bounded Errors-in-Variables Model*, SIMAX, 20 (1999), pp. 839–859.

[3] G. H. Golub and C. F. V. Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, Md, 1996.

[4] G. W. Stewart, *Introduction to Matrix Computations*, Academic Press, New York, NY, 1973.