Contents

List of Figures

List of Tables

The Devil is in the details.

0.1 Life is complex

$\mathbf{A}^*$ or $\mathbf{A}^{\mathrm{T}}$

$$\mathbb{R} \in \mathbb{C}$$

Table 0.1. *Matrix manipulations for* $\boldsymbol{A}^*$ *and* $\boldsymbol{A}^{\mathrm{T}}$.

	$\mathbf{A}^*$	$\mathbf{A}^{\mathrm{T}}$
Application	$\mathbf{A} \in \mathbb{C}^{m \times n}$ and $\mathbf{A} \in \mathbb{R}^{m \times n}$	only $\mathbf{A} \in \mathbb{R}^{m \times n}$
Fortran	`conjg(transpose(A))`	`transpose(A)`
Mathematica	`A^H` `ConjugateTranspose[A]`	`A^T` `Transpose[A]`
MATLAB	`A'` `ctranspose(A)`	`A.'` `transpose(A)`
Python	`A.T.conj()`	`A.T`

0.2 Readings

There are many excellent books available examining many facets of the least squares problem. Fuller references are in the bibliography.

Carl Freidrich Gauss
Theory of the Combination of Observations Least Subject to Errors

0.3 Least Squares

The titles are ranked by brevity.

Ilse C. F. Ipsen
Numerical Matrix Analysis: Linear System and Least Squares (128 pp)

Charles L. Lawson, and **Richard J. Hanson**
Solving Least Squares Problems (337 pp)

Åke Björk
Numerical Methods for Least Squares Problems (408 pp)

0.4 Linear Algebra and Matrix Analysis

The titles are ranked by brevity.

Alan J. Laub
Matrix Analysis for Scientists and Engineers (157 pp)

Carl D. Meyer
Matrix Analysis and Applied Linear Algebra (718 pp)

0.5 Numerical Linear Algebra

The titles are ranked by brevity.

Alan J. Laub
Computational Matrix Analysis (154 pp)

Lloyd N. Trefethen, and **David Bau, III**
Numerical Linear Algebra (361 pp)

E. Anderson, **Z. Bai**, **C. Bischof**, **S. Blackford**, **J. Demmel**, **J. Dongarra**, **J Du Croz**, **A. Greenbaum**, **S. Hammarling**, **A. McKenney**, and **D. Sorenson**
LAPACK Users' Guide (407 pp)

G. W. Stewart
Matrix Algorithms:
Volume I: Basic Decompositions (460 pp)
Volume II: Eigensystems (474 pp)

0.6 Discussions on Least Squares

Books with chapters dedicated to the topic. Sorted by author.

Gene H. Golub, **Charles F. Van Loan**
Matrix Computations, ch. 5, 6

Nicholas J. Higham
Accuracy and Stability of Numerical Algorithms, ch. 20

Cleve B. Moler
Numerical Computing with MATLAB, ch. 5

David S. Watkins
Numerical Analysis: a mathematical introduction, ch. 5

David S. Watkins
Fundamentals of Matrix Computations, ch. 3

Part I

Rudiments

Chapter 1

Least Squares Problems

1.1 Linear Systems

This story begins with the archetypal matrix–vector equation

$$\mathbf{A}x = b. \tag{1.1}$$

The matrix $\mathbf{A}$ has m rows, n columns, and has rank ρ; the vector b encodes m measurements. The solution vector x represents the n free parameters in the model. In mathematical shorthand,

$$\mathbf{A} \in \mathbb{C}_{\rho}^{m \times n}, \quad b \in \mathbb{C}^m, \quad x \in \mathbb{C}^n \tag{1.2}$$

with $\mathbb{C}$ representing the field of complex numbers. The system matrix $\mathbf{A}$ and the data vector b are given, and the task is to find the vector x.

1.1.1 $\|\mathbf{A}x - b\| = 0$

The letters in (1.1) will change, but the operation remains the same: a matrix operates on an $n-$vector and returns an $m-$vector. We can think of the matrix as a map from the space of vectors of dimension n to the space of vectors of dimension m:

$$\mathbf{A}\colon \mathbb{C}^n \mapsto \mathbb{C}^m.$$

If the vector b can be expressed a combination of the columns of the matrix $\mathbf{A}$ then there is a direct solution:

$$\mathbf{A}x = b \quad \Longrightarrow \quad x_1 a_1 + \cdots + x_n a_n = b$$

and the residual error vanishes:

$$r = \mathbf{A}x - b = \mathbf{0}$$

where the zero vector $\mathbf{0}$ is a list of m zeros. The total error, the norm of the r vector, is 0:

$$\|r\|_2^2 = \|\mathbf{A}x - b\|_2^2 = 0.$$

For the problem where the system matrix $\mathbf{A}$ is the identity matrix $\mathbf{I}_2$:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix},$$

the solution is

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix};$$

there is no residual error

$$r = \mathbf{A}x - b = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

1.1.2 $\|\mathbf{A}x - b\| > 0$

But what happens when the vector b is *not* in the column space of the matrix $\mathbf{A}$? The solution criteria must relax. Instead of seeking zero residual error, seek the least residual error

$$\|r\| = \|\mathbf{A}x - b\|.$$

Instead of a perfect solution, ask for the best solution. One such class of solutions are least squares solutions.

1.2 Least Squares Solutions

In the approximation problems which follow, the goal is to minimize the residual error and this work explores the minimal solutions under the 2−norm, the familiar norm of Pythagorus:

$$\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right\|_2 = \sqrt{x_1^2 + x_2^2}.$$

Let's construct a sample problem with $\|\mathbf{A}x - b\| > 0$ by modifying the previous example:

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}. \tag{1.3}$$

When $b_2 \neq 0$ there is no exact solution; no solution where $\|\mathbf{A}x - b\| = 0$. Consider the solutions given by

$$x_* = \begin{bmatrix} x_1 \\ 0 \end{bmatrix}; \tag{1.4}$$

the error is $\mathbf{A}x_* - b = -\begin{bmatrix} 0 \\ b_2 \end{bmatrix}$ which has a norm

$$\|r\|_2 = \|\mathbf{A}x_* - b\|_2 = |b_2|, \tag{1.5}$$

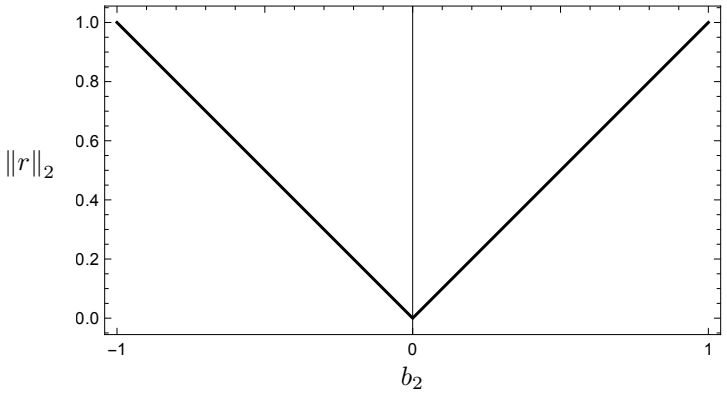


Figure 1.1. *The residual error* $\|r\|_2$ *given in* (1.5).

plotted in figure 1.1. This figure shows the solution is (1.4), the least possible error for the problem as the value of b_2 varies. In this light, the transition from an exact solution to an inexact solution is continuous.

Think of the solutions to the linear system

$$\mathbf{A}x = b$$

as being described by the inequality

$$\|\mathbf{A}x - b\|_2 \geq 0.$$

In some cases the equality is attained.

Least squares solutions are classified by the interpretation of the output. In the first case, *zonal approximation*, the output represents data at a physical zone, a point or a region. In the second case, *modal approximation*, the output represents an amplitude, a contribution for a mode. Basic examples follow.

1.2.1 Zonal Approximation

Consider the vector field F described by the gradient of a scalar field ϕ.

$$F = \nabla\phi$$

The forward problem involves taking a given potential ϕ and computing the forces F. We are instead interested in the inverse problem: measure the forces F and compute the potential *phi*.

Zonal Problem

The input and approximation for the inverse problem are depicted in 1.2. The physical field is $\phi(x)$, $0 \leq x \leq 2$, the approximation is φ_{x_k}, $k = 0, 1, 2$. For a cleaner presentation let $\varphi_{x_k} \to \varphi_k$. The first measurement x_1 represents the potential

change between $\phi(0)$ and $\phi(1)$; the second measurement x_2 the change between $\phi(1)$ and $\phi(2)$.

$$\varphi_1 - \varphi_0 \approx \delta_1$$
$$\varphi_2 - \varphi_1 \approx \delta_2$$

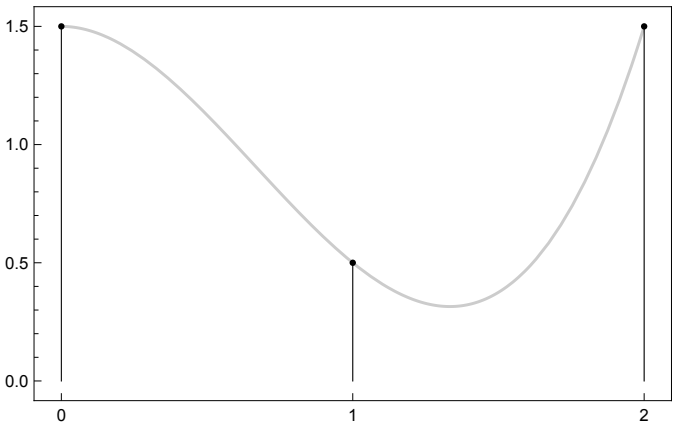


Figure 1.2. *Scalar function $\phi(x)$ (curve) and approximation φ_k (sticks).*

The system matrix

$$\mathbf{A} = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \in \mathbb{R}_2^{2\times 3}.$$

There are $m = 2$ measurements, $n = 3$ measurement locations, and the matrix rank is $\rho = 2$. Because the rank is less than the number of columns, $\rho < n$, this problem is *underdetermined*.

The linear system is

$$\begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} \varphi_0 \\ \varphi_1 \\ \varphi_2 \end{bmatrix} = \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix}. \tag{1.6}$$

Zonal Solution

The solutions for the linear system in (1.6) which minimize $\|\mathbf{A}x - b\|_2$ are

$$\begin{bmatrix} \varphi_0 \\ \varphi_1 \\ \varphi_2 \end{bmatrix} = \begin{bmatrix} -2 & -1 \\ 1 & -1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} + \gamma \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \qquad \gamma \in \mathbb{C}.$$

The color blue represents vectors in the range space, red vectors in the null space. In this way, the fundamental spaces spring to life.

There is a continuum of solutions due to the fact that

$$\mathbf{A}\,x = \mathbf{A}\left(x + \gamma \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}\right) = \mathbf{A}x + \gamma\mathbf{A}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \mathbf{A}x.$$

A quick demonstration verifies the null space vector

$$\mathbf{A}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

1.2.2 Modal Approximation

Modal Problem

In the modal approximation, the user first selects a set of basis functions to describe measurements. Popular basis functions include orthogonal polynomials, trigonometric functions, or monomials. For example, a linear regression implies a basis set of two elements: a constant function, and a linear function. This leads to the familiar equation for a straight line:

$$y(x) = a_0 + a_1 x$$

The $n = 2$ parameters represent the intercept (a_0), and the slope (a_1); each of the m measurements represents a straight line:

$$\begin{aligned} a_0 + a_1 x_1 &= y_1 \\ &\vdots \\ a_0 + a_1 x_m &= y_m. \end{aligned}$$

The goal is to simultaneously solve the set of equations.

Modal Solution

The first step is to compose the system

$$\begin{matrix} \mathbf{A} & \alpha & = & y \end{matrix}$$
$$\begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}, \tag{1.7}$$

which can be expressed using the column vectors

$$\mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}.$$

The columns of the system matrix $\mathbf{A} = \left[\begin{array}{c|c} \mathbf{1} & x \end{array}\right]$. The solution parameters can be expressed in terms of the column vectors:

$$\begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix} = \left((\mathbf{1}^{\mathrm{T}}\mathbf{1})\,(x^{\mathrm{T}}x) - (\mathbf{1}^{\mathrm{T}}x)^2 \right)^{-1} \begin{bmatrix} x^{\mathrm{T}}x & -\mathbf{1}^{\mathrm{T}}x \\ -\mathbf{1}^{\mathrm{T}}x & \mathbf{1}^{\mathrm{T}}\mathbf{1} \end{bmatrix} \begin{bmatrix} \mathbf{1}^{\mathrm{T}}y \\ x^{\mathrm{T}}y \end{bmatrix}.$$

1.2.3 Errors

When measurements are not exact, solutions are not exact. A great beauty of the method of least squares in that the quality of the solution can be quantified. An ability to discern answers like 3.0 ± 1.0 from 3.000 ± 0.0010 is invaluable. The machinery needed to compute uncertainties will be developed in following chapters.

1.3 Least Squares Problem

Emboldened by solutions to the two basic problems, we turn attention towards formalities. Starting with a the linear system $\mathbf{A}x = b$ where the matrix $\mathbf{A} \in \mathbb{C}^{m \times n}$, the data vector $b \in \mathbb{C}^m$, the least squares solution x_{LS} is defined as the set

$$x_{LS} = \left\{ x \in \mathbb{C}^n \colon \ \|\mathbf{A}x - b\|_2^2 \ \text{is minimized} \right\}. \tag{1.8}$$

The least squares solution may be a point or it may be a hyperplane. The general solution is a combination of a particular solution (in blue) and a homogenous solution (in red):

$$\begin{aligned} x_{LS} &= \mathbf{A}^{\dagger}b + \left(\mathbf{I}_n - \mathbf{A}^{\dagger}\mathbf{A}\right)y, \qquad y \in \mathbb{C}^n \\ &= x_{\dagger} + x_{\mathcal{N}} \end{aligned} \tag{1.9}$$

where the matrix $\mathbf{A}^{\dagger}$ is the pseudoinverse matrix (§???) and y is an arbitrary vector.

Chapter 2

Least Squares Solutions

Bolstered by concrete results from the previous chapter, attention now turns to an examination of solution methods through the lens of the Fundamental Theorem.

2.1 Fundamental Theorem of Linear Algebra

The Fundamental Theorem is a beautiful and powerful statement which helps to organize techniques and results in linear algebra. It is the stage which presents the story of existence and uniqueness.

Begin with the exemplar matrix

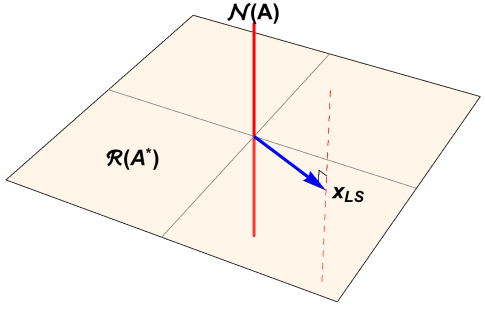
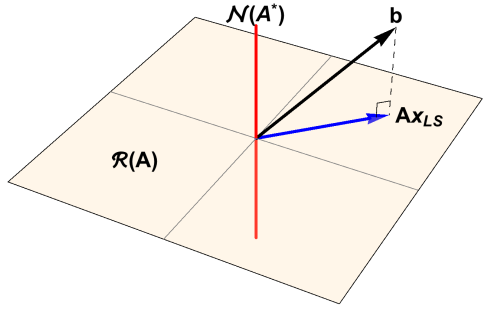
$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \in \mathbb{C}^{3\times 4}$$

which has obvious decompositions. The matrix $\mathbf{A}$ maps 4–vectors from the domain $\mathbb{C}^4$ to 3–vectors in the codomain $\mathbb{C}^3$. One way to express the Fundamental Theorem is to say that the matrix $\mathbf{A}$ induces a 4 dimensional column space and a 3 dimensional column space. For this case, the domain can be described as

$$\mathbb{C}^4 = \text{sp}\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \right\}$$

There are many statements of the Fundamental Theorem. One way describes the decomposition of the domain and codomain into the four fundamental subspaces:

Table 2.1. *The Fundamental Theorem of Linear Algebra and Least Squares for* $\boldsymbol{A} \in \mathbb{C}^{m\times n}$

Domain	Mapping	Codomain
	$\mathbf{A}\colon \mathbb{C}^n \mapsto \mathbb{C}^m$ $\mathbb{C}^n \mapsfrom \mathbb{C}^m\colon \mathbf{A}^*$	
$\mathbb{C}^n = \mathcal{R}(\mathbf{A}^*) \oplus \mathcal{N}(\mathbf{A})$		$\mathbb{C}^m = \mathcal{R}(\mathbf{A}) \oplus \mathcal{N}(\mathbf{A}^*)$

Formulaically one can write Each domain space, domain and codomain, are

Table 2.2. *The Fundamental Theorem of Linear Algebra for* $\boldsymbol{A} \in \mathbb{C}^{m\times n}$

$$\begin{array}{rccccc} \text{domain:} & \mathbb{C}^n & = & \mathcal{R}(\mathbf{A}^*) & \oplus & \mathcal{N}(\mathbf{A}) \\ \text{codomain:} & \mathbb{C}^m & = & \mathcal{R}(\mathbf{A}) & \oplus & \mathcal{N}(\mathbf{A}^*) \end{array}$$

expressed as an orthogonal decomposition of a range space and a null space.

Table 2.3. *Dimensions of the fundamental subspaces for* $\boldsymbol{A} \in \mathbb{C}_{\rho}^{m\times n}$.

$$\begin{array}{rclrcl} \dim\left(\mathcal{R}(\mathbf{A})\right) & = & \rho & \dim\left(\mathcal{N}(\mathbf{A}^*)\right) & = & m-\rho \\ \dim\left(\mathcal{R}(\mathbf{A}^*)\right) & = & \rho & \dim\left(\mathcal{N}(\mathbf{A})\right) & = & n-\rho \end{array}$$

Example from (1.6)

$$\begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}$$

Begin with the column space, also known as the solution space because the solution parameters will inhabit this space:

$$\text{Column vectors} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

The column vectors are a complete span of $\mathbb{C}^2$, and the null space is trivial:

$$\mathbb{C}^2 = \text{sp}\left\{ \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\} \oplus \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

A minimal span for $\mathbb{C}^2$ is

$$\mathbb{C}^2 = \text{sp}\left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}.$$

Next up is the row space, or the measurement space

$$\text{Row vectors} = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}$$

$$\mathbb{C}^3 = \text{sp}\left\{ \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix} \right\} \oplus \text{sp}\left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right\}$$

2.2 Singular Value Decomposition - I

The Fundamental Theorem describes the world as a orthogonal decomposition of the domain and codomain. Why not ask for an orthonormal decomposition? This is precisely what we get from the singular value decomposition.

2.2.1 SVD Theorem

Given a matrix $\mathbf{A} \in \mathbb{C}_{\rho}^{m \times n}$, a matrix with complex entries with m rows, n columns, and matrix rank $0 < \rho \leq \min(m, n)$, then there exists a decomposition of the form

$$\mathbf{A} = \mathbf{U}\,\Sigma\,\mathbf{V}^{*}$$

where

1. column vectors of unitary matrix $\mathbf{V} \in \mathbb{C}^{n \times n}$ represent an orthonormal span of the domain,
2. column vectors of unitary matrix $\mathbf{U} \in \mathbb{C}^{m \times m}$ represent an orthonormal span of the codomain,
3. Diagonal entries of $\Sigma \in \mathbb{R}^{m \times n}$ contain the singular values; the ordered, nonzero eigenvalues of the product matrix $\mathbf{A}^{*}\mathbf{A}$.

In block form

$$\mathbf{A} = \mathbf{U}\,\Sigma\,\mathbf{V}^{*} = \left[\begin{array}{c|c} \mathbf{U}_{\mathcal{R}} & \mathbf{U}_{\mathcal{N}} \end{array}\right] \left[\begin{array}{c|c} \mathbf{S} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right] \left[\begin{array}{c} \mathbf{V}_{\mathcal{R}}^{*} \\ \hline \mathbf{V}_{\mathcal{N}}^{*} \end{array}\right]$$

Column vectors span the subspaces:

$$\mathbf{V} = \left[\begin{array}{ccc|ccc} v_1 & \dots & v_{\rho} & v_{\rho+1} & \dots & v_n \end{array}\right],$$
$$\mathbf{U} = \left[\begin{array}{ccc|ccc} u_1 & \dots & u_{\rho} & u_{\rho+1} & \dots & u_m \end{array}\right].$$

$$\mathbf{U} \in \mathbb{C}^{m \times m},$$
$$\mathbf{V} \in \mathbb{C}^{n \times n},$$
$$\Sigma \in \mathbb{R}^{m \times n}.$$

Table 2.4. *Orthonormal spans for the invariant subspaces.*

		domain			codomain
$\mathcal{R}(\mathbf{A}^{*})$	$=$	$\text{span}\{v_1, \dots, v_{\rho}\}$	$\mathcal{R}(\mathbf{A})$	$=$	$\text{span}\{u_1, \dots, u_{\rho}\}$
$\mathcal{N}(\mathbf{A})$	$=$	$\text{span}\{v_{\rho+1}, \dots, v_n\}$	$\mathcal{N}(\mathbf{A}^{*})$	$=$	$\text{span}\{u_{\rho+1}, \dots, u_m\}$

$$u_j \cdot u_k = \delta_{jk},$$
$$v_j \cdot v_k = \delta_{jk}.$$

Decomposition for (1.6):

$$\begin{array}{ccccc} \mathbf{A} & = & \mathbf{U} & \Sigma & \mathbf{V}^* \end{array}$$

$$\begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \left[\begin{array}{cc|c} \sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \end{array}\right] \begin{bmatrix} \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix}$$

$$\mathbf{S} = \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \end{bmatrix}$$

2.2.2 SVD and Least Squares

A direct implication of the singular value decomposition is the homogeneous solution.

Unitary transformation

The definition of the least squares problem in (1.8) shows that the target of minimization is the quantity

$$r^* r = r^2 = \|\mathbf{A}x - b\|_2^2 .$$

One minimization strategy invokes a unitary transformation to create a simpler problem:

$$\|\mathbf{A}x - b\|_2^2 = \|\mathbf{U}^* (\mathbf{A}x - b)\|_2^2 . \tag{2.1}$$

This remarkable insight opens a door to solution. Rearranging the singular value decomposition

$$\mathbf{U}^*\mathbf{A} = \Sigma\, \mathbf{V}^*,$$

and using the block form in (2.2.1) leads to

$$\|\mathbf{A}x - b\|_2^2 = \|\Sigma\, \mathbf{V}^* x - \mathbf{U}^* b\|_2^2 = \left\| \left[\begin{array}{c|c} \mathbf{S} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right] \left[\begin{array}{c} \mathbf{V}_{\mathcal{R}}^* \\ \hline \mathbf{V}_{\mathcal{N}}^* \end{array}\right] x - \left[\begin{array}{c} \mathbf{U}_{\mathcal{R}}^* \\ \hline \mathbf{U}_{\mathcal{N}}^* \end{array}\right] b \right\|_2^2$$
$$= \left\| \left[\begin{array}{c} \mathbf{S}\mathbf{V}_{\mathcal{R}}^* x \\ \hline \mathbf{0} \end{array}\right] - \left[\begin{array}{c} \mathbf{U}_{\mathcal{R}}^* b \\ \hline \mathbf{U}_{\mathcal{N}}^* b \end{array}\right] \right\|_2^2 .$$

The range space components are now untangled from the null space components.

Pseudoinverse solution

Using the Pythagorean theorem to isolate the range and null space components of the total error for the least squares problem

$$\|\mathbf{A}x - b\|_2^2 = \underbrace{\|\mathbf{S}\mathbf{V}_{\mathcal{R}}^* x - \mathbf{U}_{\mathcal{R}}^* b\|_2^2}_{\substack{x \text{ dependence} \\ \text{under control}}} + \underbrace{\|\mathbf{U}_{\mathcal{N}}^* b\|_2^2}_{\substack{\text{residual} \\ \text{no control}}}$$

There are now two terms; the first depends upon the solution vector x, the second does not. We only have control over the first term. To minimize the total error we must drive the first term to zero. Then the total error will be given by the residual error term. The error term that is controlled by the solution vector x is this

$$\mathbf{S}\mathbf{V}_{\mathcal{R}}^* x - \mathbf{U}_{\mathcal{R}}^* b. \tag{2.2}$$

Choosing the vector x which forces this term to zero leads to the SVD solution for the least squares problem:

$$x_\dagger = \mathbf{V}_{\mathcal{R}}\mathbf{S}^{-1}\mathbf{U}_{\mathcal{R}}^* b.$$

This is also the pseudoinverse solution

$$x_\dagger = \mathbf{A}^\dagger b$$

where the (thin) pseudoinverse is

$$\mathbf{A}^\dagger = \mathbf{V}_{\mathcal{R}}\mathbf{S}^{-1}\mathbf{U}_{\mathcal{R}}^*.$$

The error that can be controlled is forced to 0; but this leaves an error which cannot be removed, a residual error defined as

$$r^2 = \|\mathbf{U}_{\mathcal{N}}^* b\|_2^2 .$$

The usually silent null space term can be heard as it pronounces the value of the total error.

To recap, the singular value decomposition leads immediately to the pseudoinverse solution and residual error.

In retrospect

Decompose the data vector in range and null space components:

$$b = b_{\mathcal{R}} + b_{\mathcal{N}}$$

(An example will be shown in (4.2.1).) Because the vector $b_{\mathcal{R}} \in \mathcal{R}(\mathbf{A})$, there exists a vector x such that $\mathbf{A}x = b_{\mathcal{R}}$.

$$\|\mathbf{A}x - b\|_2^2 = \|\mathbf{A}x - b_{\mathcal{R}} - b_{\mathcal{N}}\|_2^2 = \|b_{\mathcal{N}}\|_2^2$$

Again, the error that cannot be removed is the residual error

$$\|b_{\mathcal{N}}\|_2^2$$

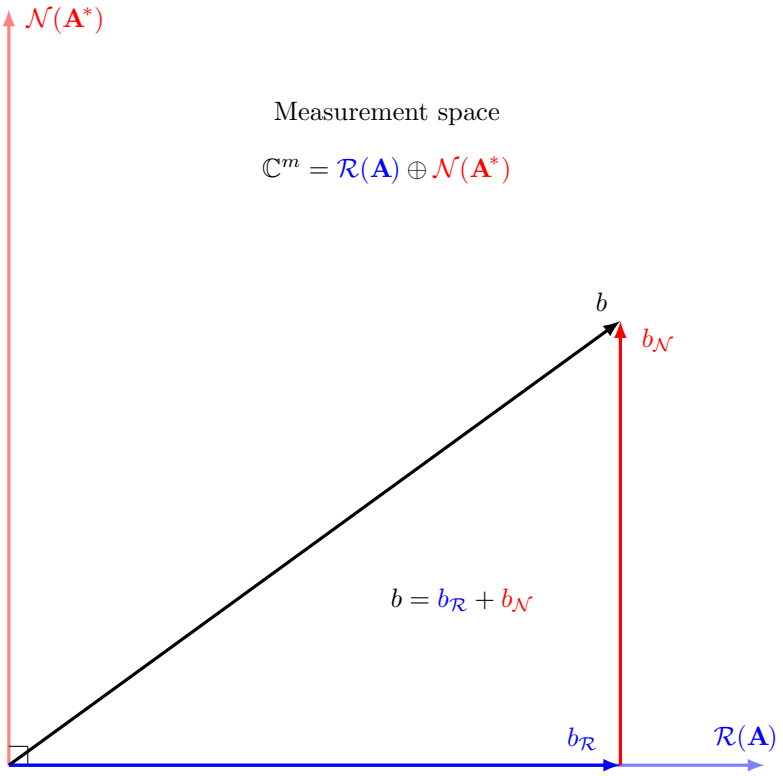


Figure 2.1. *Decomposing the data vector.*

What we shown is that the vector x which minimizes the least squares error in (**??**) is exactly the same vector given by the SVD solution in equation (2.2.2). Using a unitary transform we were able to convert the general least squares problem into a form amenable to solution using the singular value decomposition.

For the overdetermined case as we have here the usually silent null space term can be heard as it pronounces the value of the total error

$$r^2 = \|\mathbf{U}_{\mathcal{N}}^* b\|_2^2 = (\mathbf{U}_{\mathcal{N}}^* b)^* (\mathbf{U}_{\mathcal{N}}^* b) = b^* (\mathbf{U}_{\mathcal{N}} \mathbf{U}_{\mathcal{N}}^*) b \tag{2.3}$$

2.3 Singular Value Decomposition - II

2.3.1 Σ Gymnastics

Success in manipulating the singular value decomposition includes success in manipulating the Σ matrices. Think of the Σ matrix as a sabot matrix for the matrix of singular values $\mathbf{S}$. The breakdown of block diagrams: the matrix Σ $(m \times n)$ and

the matrix Σ^{T} $(n \times m)$ have different shapes, but equivalent block diagrams.

$$\Sigma = \begin{bmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$
$$\Sigma^{\mathrm{T}} = \begin{bmatrix} \mathbf{S}^{\mathrm{T}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$
$$\Sigma^{\dagger} = \begin{bmatrix} \mathbf{S}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

$$\Sigma = \left[\begin{array}{c|c} \mathbf{S}_{\rho\times\rho} & \mathbf{0}_{\rho\times n-\rho} \\ \hline \mathbf{0}_{m-\rho\times\rho} & \mathbf{0}_{m-\rho\times n-\rho} \end{array}\right]$$

Notice the interchange of the indices m and n in the transpose:

$$\Sigma^{\mathrm{T}} = \left[\begin{array}{c|c} \mathbf{S}_{\rho\times\rho} & \mathbf{0}_{\rho\times m-\rho} \\ \hline \mathbf{0}_{n-\rho\times\rho} & \mathbf{0}_{n-\rho\times m-\rho} \end{array}\right].$$

The pseudoinverse matrix has the same dimension as the parent matrix:

$$\Sigma = \left[\begin{array}{c|c} \mathbf{S}_{\rho\times\rho} & \mathbf{0}_{\rho\times n-\rho} \\ \hline \mathbf{0}_{m-\rho\times\rho} & \mathbf{0}_{m-\rho\times n-\rho} \end{array}\right]$$

$$\Sigma\,\Sigma^{\dagger} = \mathbb{I}_{m,\rho},$$
$$\Sigma^{\dagger}\Sigma = \mathbb{I}_{n,\rho}.$$

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \Sigma^{\mathrm{T}} = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \end{bmatrix},$$
$$\Sigma^{\dagger} = \begin{bmatrix} 1/\sigma_1 & 0 & 0 & 0 \\ 0 & 1/\sigma_2 & 0 & 0 \end{bmatrix}.$$

Stencil matrices

$$\Sigma\Sigma^{\dagger} = \mathbb{I}_{4,2} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{I}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$
$$\Sigma^{\dagger}\Sigma = \mathbb{I}_{2,2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathbf{I}_2.$$

Product matrices with the transpose

$$\Sigma\Sigma^{\mathrm{T}} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{S}^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

$$\Sigma^{\mathrm{T}}\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} = \mathbf{S}^2.$$

Product matrices with the pseudoinverse

$$\Sigma\Sigma^{\dagger} = \mathbb{I}_{2,2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathbf{I}_2,$$

$$\Sigma\Sigma^{\mathrm{T}} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} = \mathbf{S}^2,$$

$$\Sigma^{\dagger}\Sigma = \mathbb{I}_{4,2} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{I}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

$$\Sigma^{\mathrm{T}}\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{S}^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

2.3.2 Fundamental Projectors

Given a matrix $\mathbf{A} \in \mathbb{C}_{\rho}^{m \times n}$, a matrix with complex entries with m rows, n columns, and matrix rank $0 < \rho \leq \min(m, n)$, then there exists a

Table 2.5. *Fundamental Projectors using the pseudoinverse.*

	Range space			Null space		
domain	$\mathbf{P}_{\mathcal{R}(\mathbf{A}^*)}$	$=$	$\mathbf{A}^{\dagger}\mathbf{A}$	$\mathbf{P}_{\mathcal{N}(\mathbf{A})}$	$=$	$\mathbf{I}_n - \mathbf{A}^{\dagger}\mathbf{A}$
codomain	$\mathbf{P}_{\mathcal{R}(\mathbf{A})}$	$=$	$\mathbf{A}\,\mathbf{A}^{\dagger}$	$\mathbf{P}_{\mathcal{N}(\mathbf{A}^*)}$	$=$	$\mathbf{I}_m - \mathbf{A}\,\mathbf{A}^{\dagger}$

Table 2.6. *Fundamental Projectors using domain matrices.*

	Range space			Null space		
domain	$\mathbf{P}_{\mathcal{R}(\mathbf{A}^*)}$	$=$	$\mathbf{V}_{\mathcal{R}}\mathbb{I}_{n,\rho}\mathbf{V}_{\mathcal{R}}^*$	$\mathbf{P}_{\mathcal{N}(\mathbf{A})}$	$=$	$\mathbf{I}_n - \mathbf{V}_{\mathcal{R}}\mathbb{I}_{n,\rho}\mathbf{V}_{\mathcal{R}}^*$
codomain	$\mathbf{P}_{\mathcal{R}(\mathbf{A})}$	$=$	$\mathbf{U}_{\mathcal{R}}\mathbb{I}_{m,\rho}\mathbf{U}_{\mathcal{R}}^*$	$\mathbf{P}_{\mathcal{N}(\mathbf{A}^*)}$	$=$	$\mathbf{I}_m - \mathbf{U}_{\mathcal{R}}\mathbb{I}_{m,\rho}\mathbf{U}_{\mathcal{R}}^*$

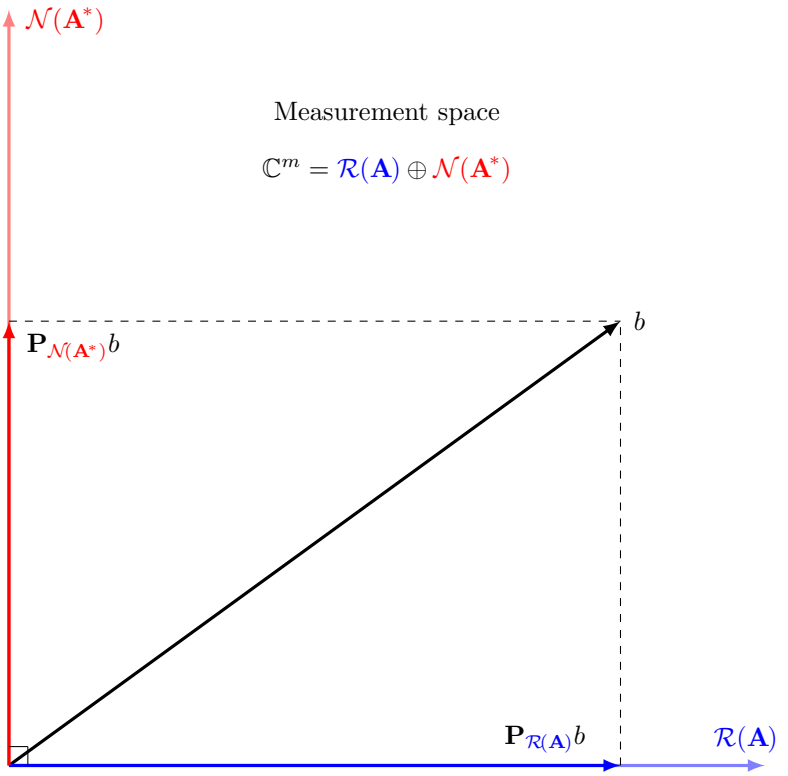


Figure 2.2. *Projections of the data vector.*

2.4 Least Squares Solution - Again

Let's revisit the canonical linear system in (1.1) the general solution in (1.9):

$$\begin{aligned} x_{LS} &= \mathbf{A}^\dagger b + \left(\mathbf{I}_n - \mathbf{A}^\dagger\mathbf{A}\right) y \\ &= \mathbf{A}^\dagger b + \mathbf{P}_{\mathcal{R}(\mathbf{A}^*)} y \end{aligned}$$

where the arbitrary vector $y \in \mathbb{C}^n$.

The projector onto the range space $\mathcal{R}(\mathbf{A}^*)$

$$\mathbf{A}^\dagger\mathbf{A} = \mathbf{V}\Sigma^\dagger\Sigma\mathbf{V}^* = \mathbf{V}_{\mathcal{R}}\mathbb{I}_{\rho,m}\mathbf{V}_{\mathcal{R}}^*$$

2.5 Reflection Invariance

Least squares solutions have a reflection invariance based upon the arithmetic fact

$$\sum r_2^2 = \sum (-r)^2 .$$

Starting with the canonical linear system in (1.1), the least squares solution is x_{LS} and the residual errors may be defined as

$$r = \mathbf{A}x - b = r \qquad \Rightarrow \qquad r = \mathbf{A}^\dagger b - b = \left(\mathbf{A}^\dagger - \mathbf{I}_m\right) T.$$

Notice that $-r = b - \mathbf{A}^\dagger b$. To reflect the data points through the solution curve us

$$\tilde{b} = b + 2r$$

$$r = b - \mathbf{A}x$$

Two equivalent ways to define the residual vector

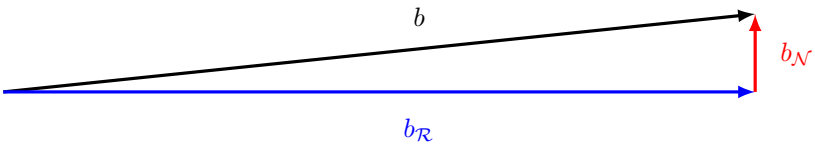


Figure 2.3. *Data vector resolved as* $b = b_{\mathcal{R}} + b_{\mathcal{N}}$.

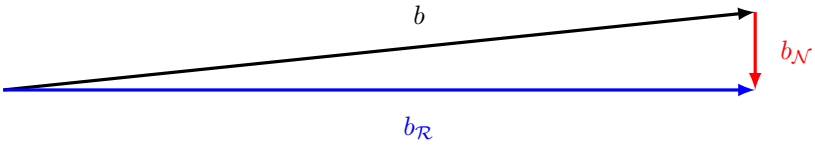


Figure 2.4. *Data vector resolved as* $b + b_{\mathcal{N}} = b_{\mathcal{R}}$.

$$x_{LS} = x_R$$
$$\left(\mathbf{A}^*\mathbf{A}\right)^{-1}\mathbf{A}^* b = \left(\mathbf{A}^*\mathbf{A}\right)^{-1}\mathbf{A}^*\left(b - 2r\right)$$

True iff $\left(\mathbf{A}^*\mathbf{A}\right)^{-1}\mathbf{A}^* r = 0$:

$$\begin{aligned}\left(\mathbf{A}^*\mathbf{A}\right)^{-1}\mathbf{A}^* r &= \left(\mathbf{A}^*\mathbf{A}\right)^{-1}\mathbf{A}^*\left(\mathbf{A}x - b\right)\\ &= \left(\left(\mathbf{A}^*\mathbf{A}\right)^{-1}\mathbf{A}^*\mathbf{A}\right)x - \left(\mathbf{A}^*\mathbf{A}\right)^{-1}\mathbf{A}^* b\\ &= x - x\\ &= \mathbf{0}.\end{aligned}$$

This test can be a filter as seen in §**??**

Table 2.7. *Reflecting the data for* $\boldsymbol{A}x - T = r$.

method	original data	reflected data
I	T	$T + 2r$
II	$\mathbf{A}^{\dagger}a - r$	$\mathbf{A}^{\dagger}a + r$

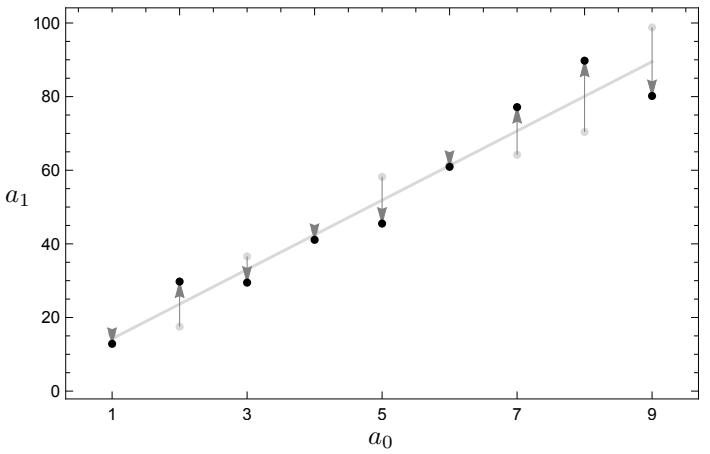


Figure 2.5. *The merit function for the learning curve showing the minimum and the value.*

Chapter 3

Modal Example I

3.1 Modal Approximation

The following example represents a problem in linear regression. A sequence of m data points (x_k, T_k), $k = 1 : m$ is are recorded. The goal is to find the best approximation to a straight line. The *trial function* is

$$T(x) = a_0 + a_1 x.$$

The residual errors are the difference between the measurements and predictions:

$$\text{residual error}_k = \underbrace{\text{measurement}_k}_{T_k} - \underbrace{\text{prediction}_k}_{T(x_k)}$$

Formally, the residual error is

$$r_k = T_k - T(x_k), \quad k = 1 : m.$$

From this springs the *merit function*, the target of minimization,

$$\begin{aligned} M(a) &= \sum_{k=1}^{m} r_k^2 \\ &= \sum_{k=1}^{m} \left(\text{measurement}_k - \text{prediction}_k\right)^2 \\ &= \sum_{k=1}^{m} \left(T_k - T\left(x_k\right)\right)^2 \\ &= \sum_{k=1}^{m} \left(T_k - a_0 - a_1 x_k\right)^2 \end{aligned} \tag{3.1}$$

The least squares solution a_{LS} is defined as

$$a_{LS} = \left\{ a \in \mathbb{C}^2 \colon \ \|T(x_k) - a_0 - a_1 x_k\|_2^2 \ \text{ is minimized} \right\}.$$

The solution satisfies

$$\nabla M(a)|_{a_{LS}} = 0. \tag{3.2}$$

3.2 Bevington Example

To provide a common reference, see the example in Bevington [**?**, ch 6]. The data is summarized below in table 3.2. The problem involves temperature measurements T_k made at position x_k on a bar in contact with two heat baths (Dirichlet boundary conditions). A conceptualization is shown in figure 3.1. Arrowheads on the bottom show the nine locations where the temperature is measured.

In the ideal linear case, the temperature at the endpoints matches the temperature of the baths, $T(x = 0 \text{ cm}) = 0°\text{C}$ and $T(x = 10 \text{ cm}) = 100°\text{C}$ which describes a line with intercept $a_0 = 0°\text{C}$ and slope $a_1 = 10°\text{C/cm}$. Such an expectation is a crude quality measure, a "sniff test".

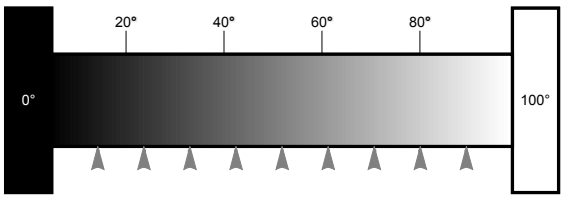


Figure 3.1. *Measuring the temperature of a bar held between two constant-temperature heat baths.*

3.2.1 Problem Statement

Muddled conceptions are wellsprings for muddled execution. Success in dealing with complicated problems in least squares flows from being able to state the problem cleanly; a good practice is to start with a table specifying the problem of interest, such as table 3.1.

The first entry is the *trial function* which defines the functional form to be applied to the data. As the name indicates, this function is an initial guess. Whether or not Nature has chosen this model remains to be seen.

The *merit function* is created by inserting the trial function into (3.1). This function is the target of minimization and can provide a crude check on the solution. Given a candidate solution a_*, compute the value of $M(a_*)$. The least squares solution has the property that $M(a_*)$ has minimum value in the neighborhood of a_*. If the solution is perturbed, one must have $M(a_*) < M(a_* + \delta a)$. When you are developing and refining least squares algorithms, you may see that the computed solution a_* changes. For overdetermined problems, the solution is unique and comparing the values of the merit function helps discriminate solutions. In a later section figure **??** will demonstrate this behavior.

The *measurements* define the quantities to be measured. It seems an obvious step, but more complex models may have ambiguities start here.

Results, or fit parameters, define the quantities to be computed using the least squares algorithm. Together with the trial function, and the measurements, we now have a clear idea of what will be measured and what will be computed.

The *residual error* specifies the difference between measurement and prediction at each point. A simple matter, apparent in the merit function, it is helpful to write it out, particularly for those who may be using the results and not intimate with the derivation.

The *system matrix* describes the measurement apparatus and contains the dependent variables. In this example we have $m = 9$ rows (measurements), $n = 2$ columns (fit parameters) and a matrix rank $\rho = 2$ (full column rank and overdetermined).

The *linear system* shows the application of the trial function to every measurement. It's a good idea to keep this image in mind.

The *ideal solution* is an infrequent visitor which helps provide a rough measure of quality. Caution is required, though. The ideal solution typically represents a concatenation of miracles which Nature may avoid. In this example, the ideal solution assumes magic barriers which absorb no heat, a bar of exact length, thermometers with exact measurements, heat baths at exact temperatures, no interaction with the local environment, etc. The hope is that systematic effects will be negligible and random effects will have 0 mean.

The next phase is to gather and record the data as shown in table 3.2. Discussion of significant digits in the input data is deferred.

3.2.2 Normal Equations via Calculus

In §6.4, Bevington solves the problem by applying calculus to the final form in (3.1), effectively solving (3.2). Introducing the notation

$$\partial_j M(a_0, a_1) = \frac{\partial M(a_0, a_1)}{\partial a_j}$$

the simultaneous equations to solve are

$$\begin{aligned} \partial_0 M(a_0, a_1) &= 0, \\ \partial_1 M(a_0, a_1) &= 0. \end{aligned} \tag{3.3}$$

These two differential equations spawn the linear system

$$\begin{aligned} -2 \sum_{k=1}^{m} (T_k - a_0 - a_1 x_k) \quad &= 0, \\ -2 \sum_{k=1}^{m} (T_k - a_0 - a_1 x_k)\, x_k &= 0. \end{aligned}$$

Distributing the summation operators creates a more revealing form

Table 3.1. *Problem statement for linear regression.*

trial function	$T(x) = a_0 + a_1 x$	°C
residual error	$r_k = T_k - a_0 - a_1 x$	°C
merit function	$M(a) = \sum_{k=1}^{m} (T_k - a_0 - a_1 x)^2$	$(°C)^2$
measurements	x_k, $k = 1{:}m$	position, cm
	T_k, $k = 1{:}m$	temperature, °C
results	$a_0 \pm \epsilon_0$	intercept, °C
	$a_1 \pm \epsilon_1$	slope, °C / cm
# of measurements	$m = 9$	rows in $\mathbf{A}$
# of parameters	$n = 2$	columns in $\mathbf{A}$
system matrix	$\mathbf{A} \in \mathbb{R}_2^{9\times 2}$	
linear system	$\begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} T_1 \\ \vdots \\ T_m \end{bmatrix}$	
ideal solution	$\begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 10 \end{bmatrix}$	
input data	table 3.2	

$$\begin{aligned} \sum_{k=1}^{m} T_k &= a_0 \sum 1 + a_1 \sum x_k, \\ \sum_{k=1}^{m} T_k x_k &= a_0 \sum x_k + a_1 \sum x_k^2, \end{aligned}$$

where summation from 1 to m is implied. (Therefore $\sum 1 = m$.) The minimization criteria are now recast as the linear system

$$\begin{bmatrix} \sum 1 & \sum x_k \\ \sum x_k & \sum x_k^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum T_k \\ \sum T_k x_k \end{bmatrix}. \tag{3.4}$$

The solution can be written immediately. Defining the determinant

$$\Delta = m \sum x_k^2 - \left(\sum x_k \right)^2, \tag{3.5}$$

the matrix inverse is

$$\begin{bmatrix} m & \sum x_k \\ \sum x_k & \sum x_k^2 \end{bmatrix}^{-1} = \Delta^{-1} \begin{bmatrix} \sum x_k^2 & -\sum x_k \\ -\sum x_k & m \end{bmatrix}. \tag{3.6}$$

Table 3.2. *Raw data and results.*

	Input		Output	
	(cm)	$(^\circ C)$	$(^\circ C)$	$(^\circ C)$
k	x_k	T_k	$T(x_k)$	r_k
1	1	15.6	14.2222	–1.37778
2	2	17.5	23.6306	6.13056
3	3	36.6	33.0389	–3.56111
4	4	43.8	42.4472	–1.35278
5	5	58.2	51.8556	–6.34444
6	6	61.6	61.2639	–0.336111
7	7	64.2	70.6722	6.47222
8	8	70.4	80.0806	9.68056
9	9	98.8	89.4889	–9.31111

The solution to equation (3.4) is the matrix product

$$\begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \Delta^{-1} \begin{bmatrix} \sum x_k^2 & -\sum x_k \\ -\sum x_k & \sum 1 \end{bmatrix} \begin{bmatrix} \sum T_k \\ \sum T_k x_k \end{bmatrix}$$

Compare the final results to Bevington's equations 6–17:

$$a_0 = \Delta^{-1} \left(\sum x_k^2 \sum T_k - \sum x_k \sum T_k x_k \right),$$
$$a_1 = \Delta^{-1} \left(m \sum T_k x_k - \sum x_k \sum T_k \right).$$

Bevington's §6–5 is a succinct explanation of error propagation. In short, measurements are inexact, therefore results will be inexact. The beauty of the method of least squares is that the error in the solution parameters can be expressed in terms of the error in the data. Measurements without uncertainties are incomplete measurements.

The computation chain begins with an estimate of the parent standard deviation which is based upon the total error:

$$s^2 \approx \frac{r^* r}{m - n}.$$

Error contributions for individual parameters are harvested from the diagonal elements of the matrix inverse in (3.6):

$$\epsilon_0^2 = \frac{r^{\mathrm{T}} r}{\Delta (m - n)} \sum x_k^2$$
$$\epsilon_1^2 = \frac{r^{\mathrm{T}} r}{\Delta (m - n)} \sum 1$$

Table 3.3. *Results for linear regression.*

fit parameters	a_0	intercept, °C
	a_1	slope, °C / cm
solution function	$T(x) = a_0 + a_1 x$	°C
solution error	$\epsilon_T^2(x) = \epsilon_0^2 + x^2\epsilon_1^2 + a_1^2\epsilon_x^2$	°C
computed solution	$\begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 4.8 \\ 9.41 \end{bmatrix} \pm \begin{bmatrix} 4.9 \\ 0.87 \end{bmatrix}$	
ideal solution	$\begin{bmatrix} \tilde{a}_0 \\ \tilde{a}_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 10 \end{bmatrix}$	
r^*r	316.6	
curvature matrix $(\mathbf{A}^*\mathbf{A})^{-1}$	$\frac{1}{180}\begin{bmatrix} 95 & -15 \\ -15 & 3 \end{bmatrix}$	
problem statement	table 3.1	
input data	table 3.2	
plots	figure 3.2	1. data and solution
	figure 3.3	2. residual errors
	figure **??**	3. merit function

3.3 Numerical Results

Results are stated in two forms. The first is an integer form free of numerical errors inherent in binary representations with finite length. This liberates one from trying to determine if errors are in the algorithm or in machine arithmetic. To aid debugging, intermediate results are also provided.

The second form represents the answer which would be provided to a customer: the fit parameters and associated errors quoted with the proper amount of significant digits.

3.3.1 Exact Form

Exact results for the fit parameters are error follow. The product matrix in (3.4) is

$$\begin{bmatrix} m & \sum x_k \\ \sum x_k & \sum x_k^2 \end{bmatrix} = \begin{bmatrix} 9 & 45 \\ 45 & 285 \end{bmatrix},$$

with determinant

$$\Delta = 540.$$

The inverse of this matrix, (3.6), is

$$\begin{bmatrix} m & \sum x_k \\ \sum x_k & \sum x_k^2 \end{bmatrix}^{-1} = \Delta^{-1} \begin{bmatrix} 285 & -45 \\ -45 & 9 \end{bmatrix}.$$

The solution vector, (3.2.2), is

$$a = \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \frac{1}{360} \begin{bmatrix} 1733 \\ 3387 \end{bmatrix}.$$

The residual error vector is

$$r = \frac{1}{360} \begin{bmatrix} -496 \\ 2207 \\ -1282 \\ -487 \\ -2284 \\ -121 \\ 2330 \\ 3485 \\ -3352 \end{bmatrix},$$

making the total error

$$r^* r = \frac{1\,139\,969}{3600}.$$

The uncertainties are then

$$\epsilon = \begin{bmatrix} \epsilon_0 \\ \epsilon_1 \end{bmatrix} = \left(360\sqrt{35}\right)^{-1} \begin{bmatrix} 108\,297\,055 \\ 3\,419\,907 \end{bmatrix}.$$

3.3.2 Computed Form

The results in the previous section are a debugging tool. This section deals with formats appropriate for formal reporting. One way to quote numbers with uncertainties is using the ± (plus – minus) notation:

$$\begin{array}{lcllll} a_0 & = & 4.8 & \pm\ 4.9 & \text{intercept} & ^\circ\text{C}, \\ a_1 & = & 9.41 & \pm\ 0.87 & \text{slope} & ^\circ\text{C / cm}. \end{array}$$

An alternative presentation uses parentheses:

$$\begin{array}{lcll} a_0 & = & 4.8\,(4.9) & \text{intercept} \quad ^\circ\text{C}, \\ a_1 & = & 9.41\,(0.87) & \text{slope} \quad ^\circ\text{C / cm}. \end{array}$$

The total error is $r^* r \approx 317$.

The uncertainty determines the number of significant digits. Common practice quotes the first two digits in the uncertainty; the location of these two digits determines the number of digits in the solution. The double precision computations are

$$\begin{aligned} \epsilon_1 &= 0.8683016476563611 \quad \text{rounded to } 0.87, \\ a_1 &= 9.408333333333333 \quad \text{rounded to } 9.41. \end{aligned}$$

At this point the model can be explored and evaluated. If the model is not acceptable, another trial function can be posed. Otherwise, the trial function becomes the solution function and is stated with error:

$$\begin{aligned} T(x) &= a_0 + a_1 x, \\ \epsilon_T^2(x) &= \epsilon_0^2 + x^2\epsilon_1^2 + a_1^2\epsilon_x^2, \end{aligned}$$

which allows for interpolation and extrapolation. What happens when the solution is extrapolated to the heat baths? The expected answers are 0°C at 0 cm, and 100°C at 10 cm:

$$\begin{aligned} T\,(0) &= (4.8 \pm 4.9)^\circ\,\mathrm{C}, \\ T\,(10) &= (99. \pm 10.)^\circ\,\mathrm{C}. \end{aligned}$$

One final thought. The method of least squares minimizes the sums of the squares of the residual errors. And in linear regression, the sum of these residuals must be 0. That is,

$$\sum_{k=1}^{m} r_k = 0.$$

This can be an quick method for evaluating solutions and data sets. Given the data and the solution parameters a a quick Python or Mathematica script can compute and sum the residuals. If a data point is omitted, the sum will not be 0. If the parameters are misquoted, the sum will not be 0. If a data point is corrupted, the sum will not be 0. Or if the solutions are for another data set, the sum will not be 0. The 0 test is simple and powerful.

Thanks to integer arithmetic, it is easy to verify that the sum of the residuals in this problem is exactly 0.

3.4 Visualization

3.4.1 Seeing the Solution

Numbers tell a story and plots bring a story to life. Besides a powerful and immediate impact, the plots are often a defense against a wide host of problems.

The first plot is the solution curve plotted against the measurements as shown in figure 3.3. The residual errors are shown as red segments which are the actual components of the residual error vector in $\mathbb{C}^9$. That is, the length of the red bars are given by the coordinates of r in the null space $\mathcal{N}(\mathbf{A}^*)$. But what is the basis for which these coordinates correspond? That will be addressed in §4.2.1.

An examination of this first plot is a first step, a crude tool which shows that the model was correctly applied. It reveals that there are no gross systematic errors; it says little about the quality of the fit. For that, more refinement is needed.

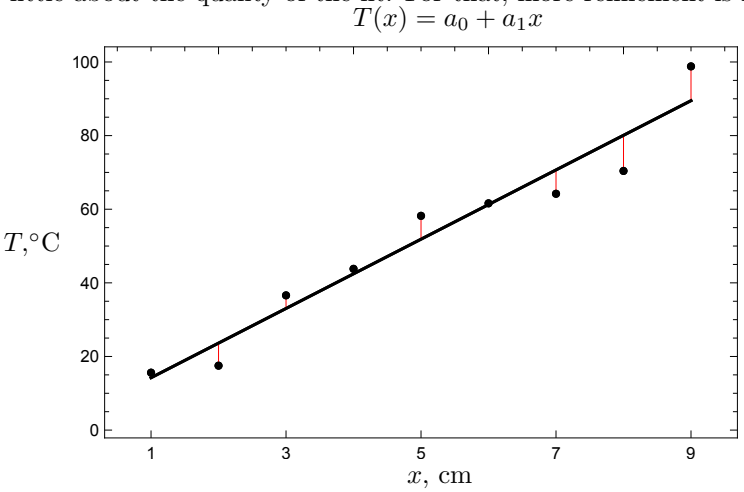


Figure 3.2. *Solution plotted against data with residual errors shown in red.*

Next, examine the residual errors as in figure 3.3. This is a more detailed examination of fit quality. The resolution increases by roughly a factor of 10 here. The dark line shows the mean of the residuals; exactly 0 in this case. The gray band represents the first standard deviation.

If data points need to be excluded, the residuals plot buttresses the case. The lone point that is five standard deviations away from the mean is quite obvious. Yes, it will be apparent on the solution plot, but that does not show rejection criteria in terms of the standard deviation.

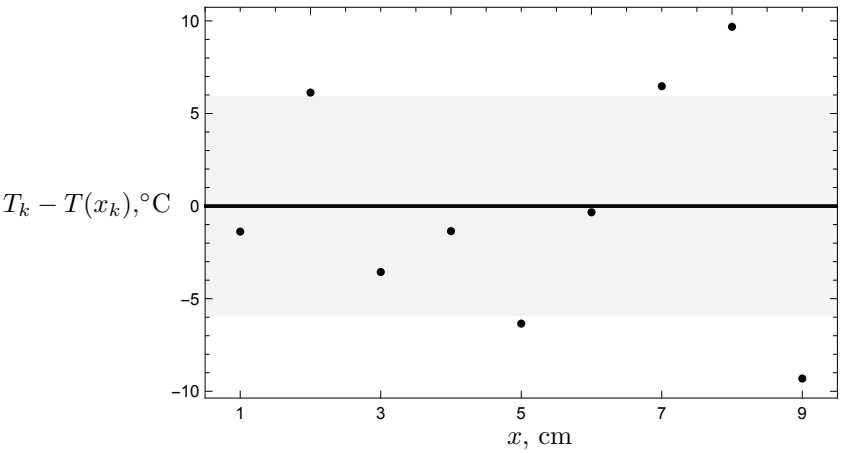


Figure 3.3. *Scatter plot of residual errors showing mean (thick line) and one standard deviation (gray box).*

Ideally, the residuals should be uncorrelated and scattered. In this case, there is a hint of correlation; a hint that the model may be extended. Or we could be a victim of low statistics – the behavior is random, but the sampling is too limited. To help think about the issue of correlation, just connect the dots as in figure 3.4. The result should be a jittery line with an oscillating appearance, not a trend line. As in this case.

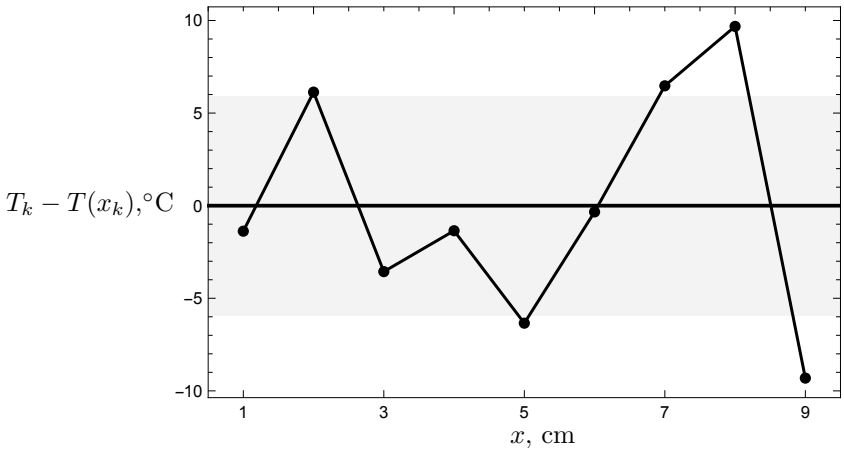


Figure 3.4. *Scatter plot of residual errors with data points connected.*

The third and final diagnostic plot, figure 3.3, puts the solution on a map of the merit function. This addresses the question of whether the solution is a true minimum and applies whether the problem is linear or nonlinear. Complicated codes can produce answers which appear valid and this plot helps to filter the good from the bad.

The plot shows a level sets of the merit function, loci where $M(a) = const.$ The least squares solution is plotted in the middle. The three concentric circles represent 1, 2, and 3 standard deviations in the fit parameters. A modified version of this plot is shown in figure 3.4.1.

The modified version explicitly labels the radii of the error ellipse, ϵ_0, and ϵ_1. The contour values are marked and create a feeling of a paraboloid with a minimum at $r^2 \approx 317$. The large arrowheads show the location of the solution precisely along the axes. The smaller escort arrowheads show the extent of the first standard deviation.

3.4.2 Seeing the Uncertainty

The interpretation of the variation of the parameters a is straightforward. A nice visualization is the "whisker plot" in figure 3.7. Given the mean a and the standard deviation ϵ, solutions can be sampled as a population. For the plot, 250 solutions a_μ were sampled and drawn against the data.

$$T_\mu(x) = a_{0_\mu} + a_{1_\mu} x, \qquad \mu = 1{:}250.$$

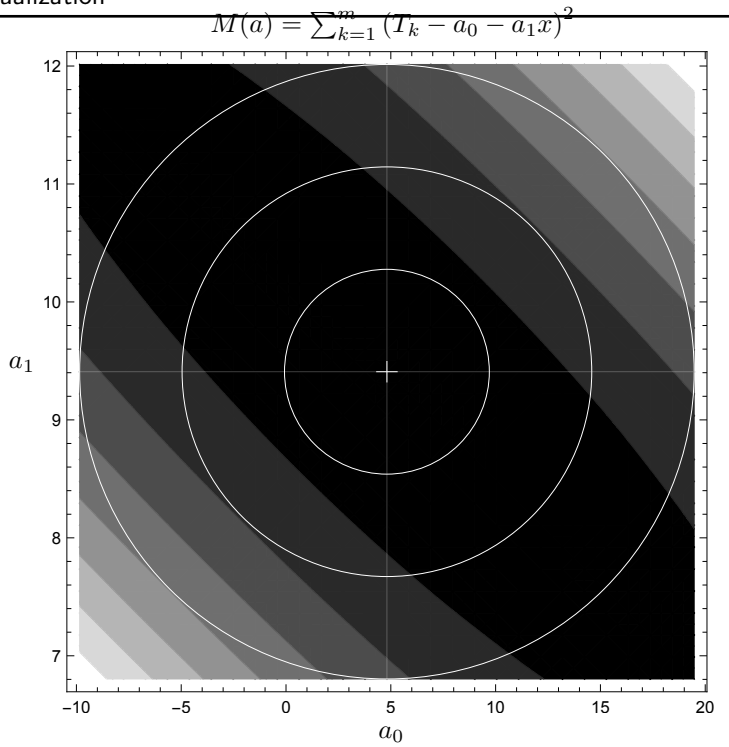


Figure 3.5. *Contour plot of merit function showing the solution (white cross) and three concentric error bands.*

Each solution represents a line. All of the lines provide a qualitative feel for the simultaneous variation of slope and intercept prescribed by the solution. Notice the white dots representing the ideal solution.

3.4.3 Digging Deeper

The whisker plot is a nice tool to quickly communicate the variation of the solution. This section describes the nuts and bolts of creating the representation. The primary tool is the assumed distribution of errors. In this exercise, it is the normal distribution:

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The curve is normalized to unity, that is,

$$\int_{-\infty}^{\infty} f(x;\mu,\sigma)\, dx = 1.$$

The mean value is μ, and the standard deviation is σ. For the data analysis, the mean represents the fit parameters a, and the standard deviation represents the error ϵ. Figure 3.4 shows these connections explicitly and provides a qualitative representation for the quality of the parameters.

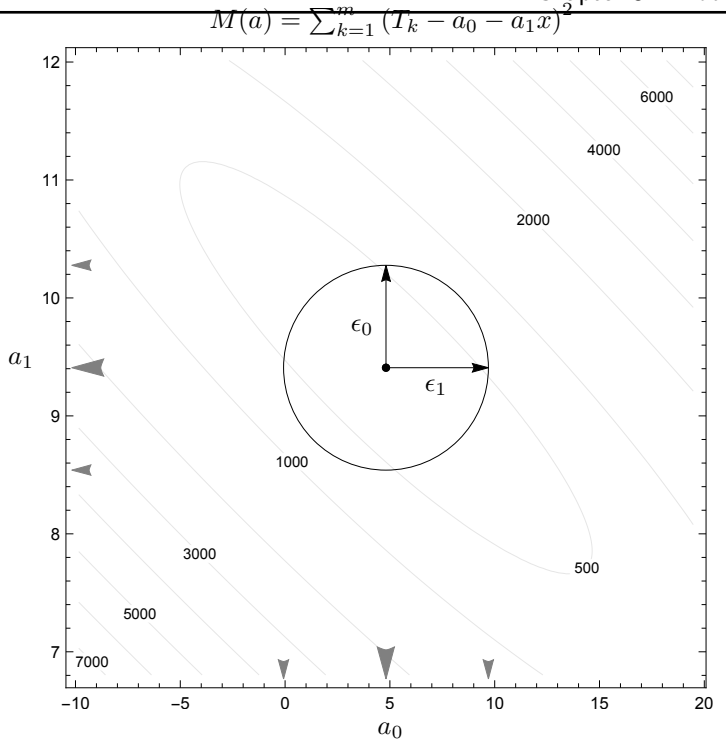


Figure 3.6. *Another look at the merit function showing the primary error ellipse.*

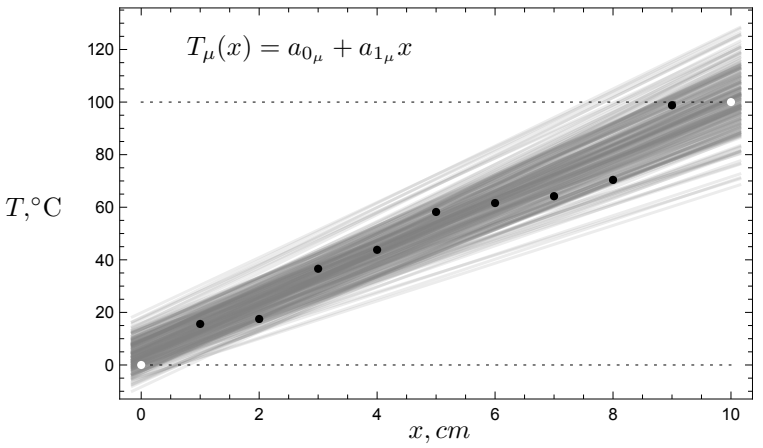
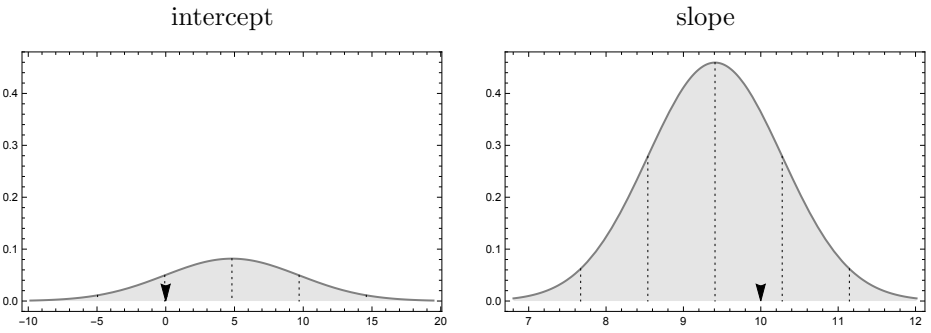


Figure 3.7. *Whisker plot showing 250 randomly sampled solutions.*

The numbers on the vertical axis are devoid of meaning and are included only to emphasize the point that the two distributions were plotted on a common scale. Knowing that, it is apparent that the slope measurement on the right is much more precise that the intercept measurement on the left.

The arrowheads locate the ideal solution and the dotted lines demarcate the

Table 3.4. *The solution parameters expressed as normal distributions.*



first three standard deviations. This provides a visual gauge of the accuracy and precision of the calculation.

Figure 3.8 shows the actual solutions used in the whisker plot as 250 points in the solution space $\mathcal{R}(\mathbf{A}^*)$. The symbol "+" marks the ideal solution.

A crude, but useful tool, is the counting of the number of points within the error bands which allows for a direct comparison with continuum theory. For example, the innermost circle should contain, in the continuum limit, 68.27% of the data points. The second band should envelope 95.45% of the points, and so on. Table 3.5 displays the actual census and compares to the continuum results.

Table 3.5. *Comparing samples to ideal normal distribution.*

annulus	count	rel. area	count density	normalized density	cumulative density	limit
1	88	1	88.0	64.83%	64.83%	68.27%
2	115	3	38.3	28.24%	93.07%	95.45%
3	41	5	8.2	6.41%	99.16%	99.73%
4	6	5	1.2	0.88%	100.0%	99.99%
	250		135.7	100.00%		

The first column identifies the annular region in which the points were counted, and the census for that region. For example, the fourth annulus contains just 6 points. The area column displays the relative area of the annular regions computed

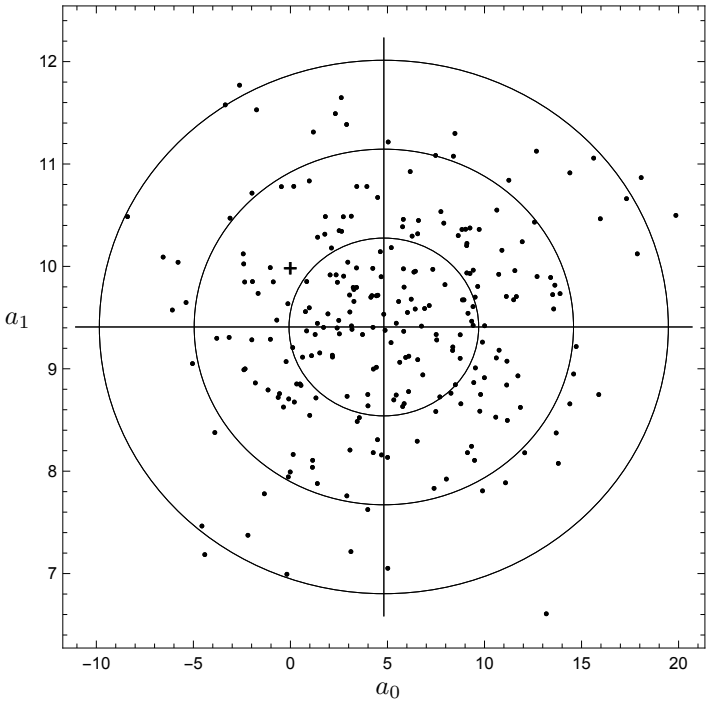


Figure 3.8. *Scatter plot showing sampling of 250 solution points in* $\mathcal{R}(\boldsymbol{A}^*)$.

using the area formula

$$\text{area}_k = \pi \left(k^2 - (k-1)^2 \right), \qquad k = 1{:}5.$$

The relative area removes the constant π; this does not affect the final answer and it simplifies the cross check. The count density divides the counts in a zone by the relative area of that zone, e.g. $115/3 = 38.3$. The normalized density is the ratio of the count density to the sum of the count densities, e.g. $38.3/135.7 = 28.24$. The cumulative density accumulates the normalized densities. The final column is the continuum limit

$$\int_{-k\sigma}^{k\sigma} f\,(x; \mu = 1, \sigma = 1)\, dx.$$

Chapter 4

Modal Example II
Other Methods

For comparison, other solution methods are applied to the Bevington data. These methods will reinforce the geometry of least squares.

4.1 Normal Equations from Vectors

In §3.2.2, the normal equations appeared from the linear system resulting from the two partial differential equations in (3.3). Another approach starts with the linear equation $\mathbf{A}x = b$ and uses column vectors. The data is posed in terms of the column vectors in table 3.2. These, plus a constant vector, are the elements of composition:

$$\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad x = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \end{bmatrix}, \quad T = \frac{1}{10} \begin{bmatrix} 156 \\ 175 \\ 366 \\ 438 \\ 582 \\ 616 \\ 642 \\ 704 \\ 988 \end{bmatrix}.$$

There are two column vectors in the system matrix $\mathbf{A}$:

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{1} & x \end{array} \right]$$

In the service of a crisp mental image, the linear system is written explicitly:

$$\begin{array}{ccccc} \mathbf{A} & & a & = & T \\ \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \\ 1 & 7 \\ 1 & 8 \\ 1 & 9 \end{bmatrix} & & \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} & = & \frac{1}{10} \begin{bmatrix} 156 \\ 175 \\ 366 \\ 438 \\ 582 \\ 616 \\ 642 \\ 704 \\ 988 \end{bmatrix} \end{array} . \tag{4.1}$$

4.1.1 Composing the normal equations

The system encodes nine linear problems:

$$\begin{aligned} a_0 + a_1 x_1 &= T_1, \\ &\vdots \\ a_0 + a_1 x_9 &= T_9. \end{aligned}$$

But there are no parameters a_0 and a_1 which solve all nine equations. More formally, the data vector T is not in the column space of $\mathbf{A}$. There are no parameters a_0 and a_1 such that

$$a_0 \mathbf{1} + a_1 x = T.$$

One strategy is to compose a solution does have a solution

$$\mathbf{A}^* \mathbf{A} a = \mathbf{A}^* T.$$

Certainly the vector $\mathbf{A}^* T$ is in the column space of $\mathbf{A}^*$ – the coordinates are T! The solution is

$$\begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = (\mathbf{A}^* \mathbf{A})^{-1} \mathbf{A}^* T.$$

For the Bevington data set the dot products are,

$$\begin{aligned} \mathbf{1}^{\mathrm{T}} \mathbf{1} &= m = 9, \\ \mathbf{1}^{\mathrm{T}} x &= x^{\mathrm{T}} \mathbf{1} = 45, \\ x^{\mathrm{T}} x &= 285, \\ \mathbf{1}^{\mathrm{T}} T &= \frac{4667}{10}, \\ x^{\mathrm{T}} T &= 2898. \end{aligned}$$

Other intermediate products include the product matrix,

$$\mathbf{A}^*\mathbf{A} = \begin{bmatrix} \mathbf{1}^{\mathrm{T}}\mathbf{1} & \mathbf{1}^{\mathrm{T}}x \\ x^{\mathrm{T}}\mathbf{1} & x^{\mathrm{T}}x \end{bmatrix} = \begin{bmatrix} 9 & 45 \\ 45 & 285 \end{bmatrix},$$

its determinant,

$$\Delta = \left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)\left(x^{\mathrm{T}}x\right) - \left(\mathbf{1}^{\mathrm{T}}x\right)^2, \tag{4.2}$$

inverse,

$$\left(\mathbf{A}^*\mathbf{A}\right)^{-1} = \Delta^{-1}\begin{bmatrix} x^{\mathrm{T}}x & -\mathbf{1}^{\mathrm{T}}\mathbf{1} \\ -\mathbf{1}^{\mathrm{T}}\mathbf{1} & \mathbf{1}^{\mathrm{T}}x \end{bmatrix}$$

and the vector,

$$\mathbf{A}^*T = \begin{bmatrix} \mathbf{1}^{\mathrm{T}}T \\ x^{\mathrm{T}}T \end{bmatrix} = \frac{1}{10}\begin{bmatrix} 4667 \\ 28\,980 \end{bmatrix}.$$

The solution is again

$$a = \Delta^{-1}\begin{bmatrix} x^{\mathrm{T}}x & -\mathbf{1}^{\mathrm{T}}\mathbf{1} \\ -\mathbf{1}^{\mathrm{T}}\mathbf{1} & \mathbf{1}^{\mathrm{T}}x \end{bmatrix}\begin{bmatrix} \mathbf{1}^{\mathrm{T}}T \\ x^{\mathrm{T}}T \end{bmatrix}$$

$$\begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \Delta^{-1}\begin{bmatrix} \left(x^{\mathrm{T}}x\right)\left(\mathbf{1}^{\mathrm{T}}T\right) - \left(\mathbf{1}^{\mathrm{T}}x\right)\left(x^{\mathrm{T}}T\right) \\ \left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)\left(x^{\mathrm{T}}T\right) - \left(\mathbf{1}^{\mathrm{T}}x\right)\left(\mathbf{1}^{\mathrm{T}}T\right) \end{bmatrix} \tag{4.3}$$

leading to the same solution presented in §3.3.1.

4.2 Singular Value Decomposition

The singular value decomposition is an $x-$ray revealing the structure of the fundamental spaces.

4.2.1 Computing the SVD

Solution steps

1. Compute $\lambda\left(\mathbf{A}^*\mathbf{A}\right)$.
2. Educated guess at domain matrix $\mathbf{V}$.
3. Compute codomain matrix $\mathbf{U}$.

The least squares problem delivers a singular value decomposition (SVD) without the muss and fuss of solving an eigensystem. The SVD is given by the matrix product

$$\mathbf{A} = \mathbf{U}\,\Sigma\,\mathbf{V}^*.$$

For the full column rank problem we have we can expand in the following block decomposition

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{U}_{\mathcal{R}} & \mathbf{U}_{\mathcal{N}} \end{array}\right] \left[\begin{array}{c} \mathbf{S} \\ \hline \mathbf{0} \end{array}\right] \left[\begin{array}{c} \mathbf{V}_{\mathcal{R}}^{*} \end{array}\right]$$

The $\mathbf{S}$ matrix contains the singular values

$$\mathbf{S} = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}.$$

The column vectors of the matrix $\mathbf{V}_{\mathcal{R}}$ represent an orthonormal basis for the row space (domain). The column vectors of the matrix $\mathbf{U}_{\mathcal{R}}$ represent two of the nine vectors in an orthonormal basis for the column space (codomain).

Singular values

The singular value spectrum of the matrix $\mathbf{A}$ is the square root of the (non-zero) eigenvalues of the product matrix $\mathbf{A}^{*}\mathbf{A}$

$$\sigma\left(\mathbf{A}\right) = \sqrt{\lambda\left(\mathbf{A}^{*}\mathbf{A}\right)}.$$

The eigenvalues of the product matrix are the roots of the characteristic polynomial $p(\lambda)$ for said matrix.

$$p(\lambda) = \lambda^2 - \lambda \operatorname{tr}\left(\mathbf{A}^{*}\mathbf{A}\right) + \det\left(\mathbf{A}^{*}\mathbf{A}\right)$$

The determinant appeared in equations (3.5) and (4.2); the trace is $\operatorname{tr}\left(\mathbf{A}^{*}\mathbf{A}\right) = \mathbf{1}^{\mathrm{T}}\mathbf{1} + x^{\mathrm{T}}x$. The singular values are then

$$\sigma = \sqrt{\frac{1}{2}\left(\mathbf{1}^{\mathrm{T}}\mathbf{1} + x^{\mathrm{T}}x \pm \sqrt{4\left(\mathbf{1}^{\mathrm{T}}x\right)^2 - \left(\mathbf{1}^{\mathrm{T}}\mathbf{1} - x^{\mathrm{T}}x\right)^2}\right)}.$$

(The astute reader will notice that the discriminant does not seem to have the familiar form of $b^2 - 4ac$. The earnest reader will discover why this is so.) The singular value spectrum for these data is

$$\sigma = \sqrt{3\left(49 \pm \sqrt{2341}\right)} \approx \left(17.0924, 1.35954\right).$$

We now have the sigma matrix and the matrix of singular values $\mathbf{S}$:

$$\Sigma = \left[\begin{array}{c} \mathbf{S} \\ \hline \mathbf{0} \end{array}\right] = \left[\begin{array}{cc} \sigma_1 & 0 \\ 0 & \sigma_2 \\ \hline 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{array}\right] \approx \left[\begin{array}{cc} 17.0924 & 0 \\ 0 & 1.35954 \\ \hline 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{array}\right].$$

Domain matrix

We can skip the eigenvector problem. To find the domain matrix we exploit the singular value decomposition of the product matrix

$$\Sigma^{\mathrm{T}}\Sigma = \begin{bmatrix} \mathbf{S} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{S} \\ \mathbf{0} \end{bmatrix} = \mathbf{S}^2$$

$$\mathbf{V}_{\mathcal{R}}\mathbf{S}^2\mathbf{V}_{\mathcal{R}}^{*} = \mathbf{A}^{*}\mathbf{A}. \tag{4.4}$$

By the singular value theorem the matrix is unitary and will be a rotation matrix, a reflection matrix or a convolution. We begin by trying a rotation matrix

$$\mathbf{V}_{\mathcal{R}} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{4.5}$$

which is colored blue because the column vectors belong to $\mathcal{R}(\mathbf{A}^{*})$. The objective is to find the angle θ. The immediate result of equations (**??**), (4.4), and (4.5) is

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \mathbf{S}^2 \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}^{\mathrm{T}} = \mathbf{A}^{*}\mathbf{A},$$

$$\begin{bmatrix} \sigma_1^2\cos^2\theta + \sigma_2^2\sin^2\theta & \left(\sigma_1^2 - \sigma_2^2\right)\cos\theta\sin\theta \\ \left(\sigma_1^2 - \sigma_2^2\right)\cos\theta\sin\theta & \sigma_2^2\cos^2\theta + \sigma_1^2\sin^2\theta \end{bmatrix} = \begin{bmatrix} \mathbf{1}^{\mathrm{T}}\mathbf{1} & \mathbf{1}^{\mathrm{T}}x \\ x^{\mathrm{T}}\mathbf{1} & x^{\mathrm{T}}x \end{bmatrix}.$$

which presents multiple solution paths for the angle θ. For example

$$\cos\theta = \sqrt{\frac{\sigma_2^2 - \mathbf{1}^{\mathrm{T}}\mathbf{1}}{\sigma_2^2 - \sigma_1^2}} = \sqrt{\frac{x^{\mathrm{T}}x - \sigma_1^2}{\sigma_2^2 - \sigma_1^2}}. \tag{4.6}$$

This implies

$$\sin\theta =$$

The domain matrix is now

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{\mathcal{R}} \end{bmatrix} = \left(\sigma_2^2 - \sigma_1^2\right)^{-\frac{1}{2}} \begin{bmatrix} \sqrt{\sigma_2^2 - \mathbf{1}^{\mathrm{T}}\mathbf{1}} & -\sqrt{\mathbf{1}^{\mathrm{T}}\mathbf{1} - \sigma_1^2} \\ \sqrt{\mathbf{1}^{\mathrm{T}}\mathbf{1} - \sigma_1^2} & \sqrt{\sigma_2^2 - \mathbf{1}^{\mathrm{T}}\mathbf{1}} \end{bmatrix}.$$

Using the data set at hand

$$\begin{bmatrix} \mathbf{V}_{\mathcal{R}} \end{bmatrix} = \begin{bmatrix} \sqrt{\frac{1}{2} - \frac{23}{\sqrt{2341}}} & -\sqrt{\frac{1}{2} + \frac{23}{\sqrt{2341}}} \\ \sqrt{\frac{1}{2} + \frac{23}{\sqrt{2341}}} & \sqrt{\frac{1}{2} - \frac{23}{\sqrt{2341}}} \end{bmatrix} \approx \begin{bmatrix} 0.156956 & -0.987606 \\ 0.987606 & 0.156956 \end{bmatrix}$$

Codomain matrix

The final component is of course the codomain matrix. Knowing the decomposition for the adjoint matrix $\mathbf{A}^*$ and that the linear system is overdetermined we can write

$$\mathbf{U}_{\mathcal{R}}^* = \mathbf{S}^{-1}\mathbf{V}_{\mathcal{R}}^*\mathbf{A}^*.$$

The kth column vector of this matrix has the compact form

$$[\mathbf{U}_{\mathcal{R}}^*]_k = \left(\sigma_1^2 - \sigma_2^2\right)^{-\frac{1}{2}} \begin{bmatrix} \sigma_1^{-2}\left(\sqrt{\sigma_2^2 - \mathbf{1}^{\mathrm{T}}\mathbf{1}} - x\sqrt{\mathbf{1}^{\mathrm{T}}\mathbf{1} - \sigma_1^2}\right) \\ \sigma_2^{-2}\left(\sqrt{\sigma_2^2 - \mathbf{1}^{\mathrm{T}}\mathbf{1}} + x\sqrt{\mathbf{1}^{\mathrm{T}}\mathbf{1} - \sigma_1^2}\right) \end{bmatrix}.$$

Using the following shorthand,

$$f(x, y) = \sqrt{x + y/\sqrt{2341}},$$

the range space component of the codomain matrix can be written as

$$\mathbf{U}_{\mathcal{R}} = \left(6\sqrt{10}\right)^{-1} \begin{bmatrix} f(68, -3212) & -f(68, 3212) \\ f(47, -2003) & -f(47, 2003) \\ f(32, -968) & -f(32, 968) \\ f(23, -107) & -f(23, 107) \\ f(50, 580) & -f(50, -580) \\ f(23, 1093) & -f(23, -1093) \\ f(32, 1432) & f(32, -1432) \\ f(47, 1597) & f(47, -1597) \\ f(68, 1588) & f(68, -1588) \end{bmatrix}$$

If one wishes to complete the codomain matrix, use the Gram-Schmidt orthonormalization process on the matrix

$$\mathbf{U} = \begin{bmatrix} [\mathbf{U}_{\mathcal{R}}]_1 & [\mathbf{U}_{\mathcal{R}}]_2 & e_{1,9} & e_{2,9} & e_{3,9} & e_{4,9} & e_{5,9} & e_{6,9} & e_{7,9} \end{bmatrix}$$

starting with column three.

Completing the codomain matrix with an orthonormal span of the null space is optional and can be done by feeding the range space components and a complementary set of unit vectors into a Gram-Schmidt algorithm.

Error terms

$$\epsilon_k^2 = \frac{(\mathbf{A}\alpha - T)^* (\mathbf{A}\alpha - T)}{m - n} \left[(\mathbf{A}^*\mathbf{A})^{-1}\right]_{kk}$$

The error terms can be computed after this step. Given that the product matrix decomposition in (4.4) is a singular value decomposition we can trivially

write the inverse matrix as

$$(\mathbf{A}^*\mathbf{A})^{-1} = \mathbf{V}_{\mathcal{R}}\mathbf{S}^{-2}\mathbf{V}_{\mathcal{R}}^* = \frac{1}{180}\begin{bmatrix} 95 & -15 \\ -15 & 3 \end{bmatrix}$$

The error terms in (**??**) become

$$\begin{aligned}\epsilon &= \sqrt{\frac{r^*r}{\mathbf{1}^{\mathrm{T}}\mathbf{1}-n}}\sqrt{\frac{1}{\sigma_1\sigma_2}}\sqrt{\begin{bmatrix} \frac{\cos^2\theta}{\sigma_2^2}+\frac{\sin^2\theta}{\sigma_1^2} \\ \frac{\cos^2\theta}{\sigma_1^2}+\frac{\sin^2\theta}{\sigma_2^2} \end{bmatrix}}, \\ &= \sqrt{\frac{r^*r}{m-n}}\sqrt{\begin{bmatrix} \sigma_1^2-\sqrt{(\sigma_2^2-\mathbf{1}^{\mathrm{T}}\mathbf{1})(\sigma_2^2-\sigma_1^2)} \\ \sigma_2^2+\sqrt{(\sigma_2^2-\mathbf{1}^{\mathrm{T}}\mathbf{1})(\sigma_2^2-\sigma_1^2)} \end{bmatrix}}.\end{aligned}$$

To close this section, look at a numeric representation of the $\mathbf{U}$ matrix:

$$\mathbf{U} = \begin{bmatrix} 0.0670 & -0.611 & 0.656 & -0.0132 & -0.0780 & -0.137 & -0.193 & -0.240 & -0.270 \\ 0.125 & -0.496 & -0.749 & -0.0927 & -0.113 & -0.147 & -0.184 & -0.215 & -0.233 \\ 0.183 & -0.380 & 0 & 0 & 0 & 0 & 0 & 0 & 0.907 \\ 0.240 & -0.265 & 0 & 0 & 0 & 0 & 0 & 0.920 & -0.159 \\ 0.298 & -0.149 & 0 & 0 & 0 & 0 & 0.924 & -0.142 & -0.123 \\ 0.356 & -0.0337 & 0 & 0 & 0 & 0.910 & -0.150 & -0.117 & -0.0858 \\ 0.414 & 0.0817 & 0 & 0 & 0.867 & -0.198 & -0.141 & -0.0930 & -0.0490 \\ 0.471 & 0.197 & 0 & 0.755 & -0.321 & -0.209 & -0.132 & -0.0685 & -0.0123 \\ 0.529 & 0.313 & 0.0937 & -0.649 & -0.356 & -0.219 & -0.124 & -0.0441 & 0.0245 \end{bmatrix}$$

Visualization

With the singular value decomposition in hand, the domain space plots become more concrete and we do so below beginning in figure (4.2). The black vector represents the measurements

$$T = \mathbf{A}a - \mathbf{R}$$

$$y \in \mathbb{C}^9, \quad \mathbf{A}a \in \mathcal{R}(\mathbf{A}) \subseteq \mathbb{C}^2, \quad r \in \mathcal{N}(\mathbf{A}^*) \subseteq \mathbb{C}^7$$

$$T = \frac{1}{10}\begin{bmatrix} 156 \\ 175 \\ 366 \\ 438 \\ 582 \\ 616 \\ 642 \\ 704 \\ 988 \end{bmatrix}$$

The closest point to the data vector in the range $\mathcal{R}(\mathbf{A})$ is

$$\mathbf{A}\,a = \frac{1}{360}\begin{bmatrix} 5120 \\ 8507 \\ 11\,894 \\ 15\,281 \\ 18\,668 \\ 22\,055 \\ 25\,442 \\ 28\,829 \\ 32\,216 \end{bmatrix} = \alpha_1[\mathbf{U}_{\mathcal{R}}]_1 + \alpha_2[\mathbf{U}_{\mathcal{R}}]_2 \;\in \mathcal{R}(\mathbf{A})$$

where the coordinates are

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \left(30\left(\sqrt{2341}-31\right)\sqrt{4682}+58\sqrt{2341}\right)^{-1} \times \begin{bmatrix} 4\,104\,889+75\,341\sqrt{2341} \\ 3\sqrt{15}\left(753\,593-15\,933\sqrt{2341}\right) \end{bmatrix}$$

$$\approx \begin{bmatrix} 171.733 \\ -4.45594 \end{bmatrix}$$

$$T_{\mathcal{R}} \approx 171.733u_1 - 4.45594u_2$$

Another way to think of the geometry is to pose the merit function in terms of the range space vectors

$$M(\alpha) = \|T - \alpha_1 u_1 - \alpha_2 u_2\|_2^2 .$$

The minimizer is given by (4.2.1) and displayed in figure 4.1.

$$M(171.733, -4.45594) \approx 17.7949.$$

The residual error vector lies entirely in the null space $\mathcal{N}(\mathbf{A}^*)$

$$r = -\frac{1}{360}\begin{bmatrix} 496 \\ -2207 \\ 1282 \\ 487 \\ 2284 \\ 121 \\ -2330 \\ -3485 \\ 3352 \end{bmatrix}$$

$$= \alpha_3[\mathbf{U}_{\mathcal{N}}]_1 + \alpha_4[\mathbf{U}_{\mathcal{N}}]_2 + \alpha_5[\mathbf{U}_{\mathcal{N}}]_3 + \alpha_6[\mathbf{U}_{\mathcal{N}}]_4 + \alpha_7[\mathbf{U}_{\mathcal{N}}]_5 + \alpha_8[\mathbf{U}_{\mathcal{N}}]_6 + \alpha_9[\mathbf{U}_{\mathcal{N}}]_7$$
$$\in \mathcal{N}(\mathbf{A}^*)$$

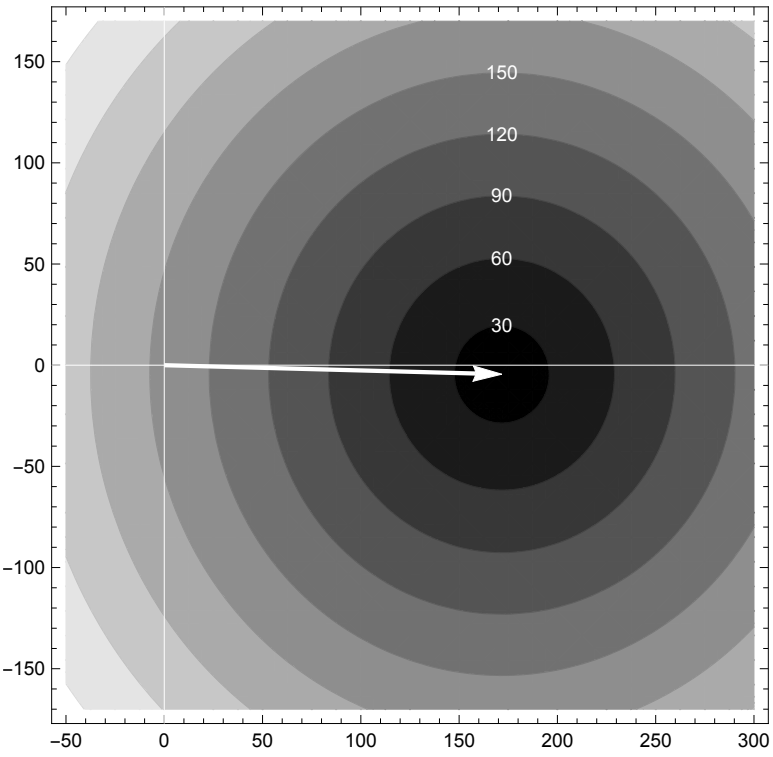


Figure 4.1. *The solution vector (white arrow) is the mixture of u_1 and u_2 which eliminates the error.*

The coordinates are now

$$\begin{bmatrix} \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \\ \alpha_7 \\ \alpha_8 \\ \alpha_9 \end{bmatrix} = \left(60\sqrt{24\,747\,709}\right)^{-1} \begin{bmatrix} 680\sqrt{7\,815\,066} \\ -1933\sqrt{3\,907\,533} \\ -3621\sqrt{186\,073} \\ 6679\sqrt{10\,434} \\ 13\,406\sqrt{29\,526} \\ 2196\sqrt{85\,386} \\ 641\sqrt{3\,344\,285} \end{bmatrix}$$

$$T = \mathbf{U}\alpha = \mathbf{A}a - r$$

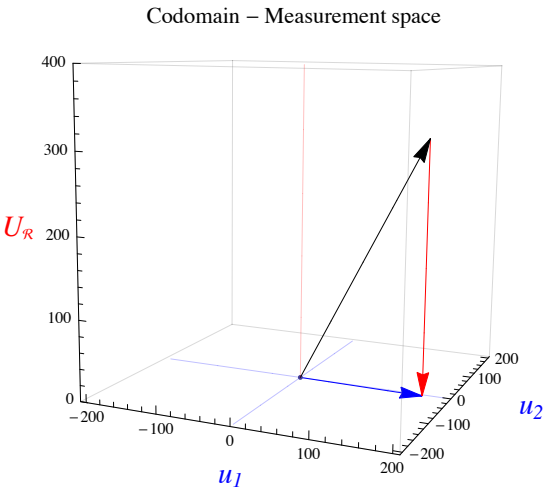


Figure 4.2. *Measurement space $\mathcal{R}(\boldsymbol{A})$ for the Bevington example.*

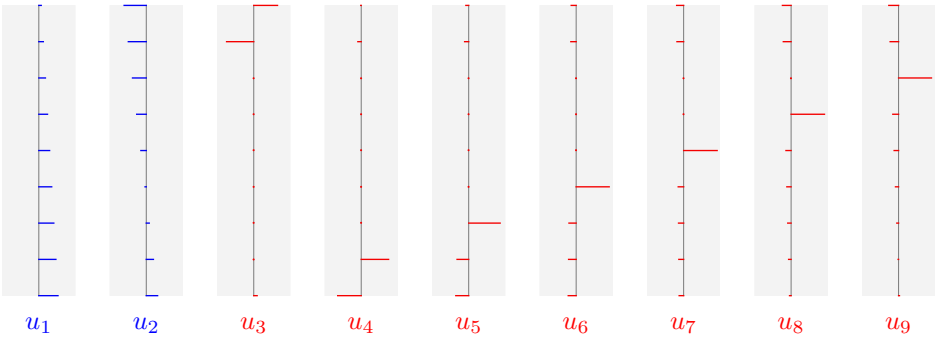


Table 4.1. *The column vectors of* $\mathbf{U}$. *The gray box represents the maximum length an element may have:* $[-1, 1]$.

The decomposition shown in figure 2.1.

$$
\begin{array}{ccccccc}
T & = & T_{\mathcal{R}} & + & T_{\mathcal{N}} \\
\frac{1}{360}\begin{bmatrix} 5616 \\ 6300 \\ 13\,176 \\ 15\,768 \\ 20\,952 \\ 22\,176 \\ 23\,112 \\ 25\,344 \\ 35\,568 \end{bmatrix} & = & \frac{1}{360}\begin{bmatrix} 5120 \\ 8507 \\ 11\,894 \\ 15\,281 \\ 18\,668 \\ 22\,055 \\ 25\,442 \\ 28\,829 \\ 32\,216 \end{bmatrix} & + & \frac{1}{360}\begin{bmatrix} 496 \\ -2207 \\ 1282 \\ 487 \\ 2284 \\ 121 \\ -2330 \\ -3485 \\ 3352 \end{bmatrix}
\end{array}
$$

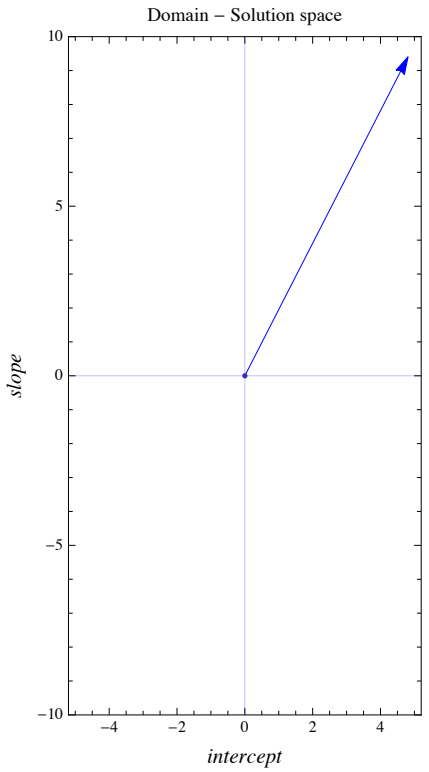


Figure 4.3. *Minimization occurs in the codomain.*

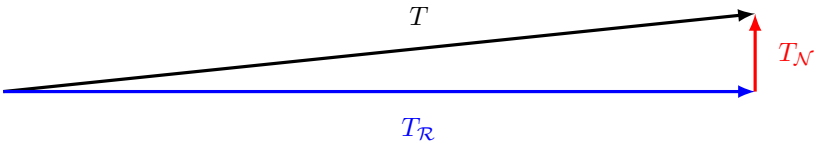


Figure 4.4. *Data vector resolved into range and null space components.*

Figure 4.5. *Decomposing* $\|r = T_{\mathcal{N}}\|_2^2$ *into residual error terms* r_k^2 *of table* **??**.

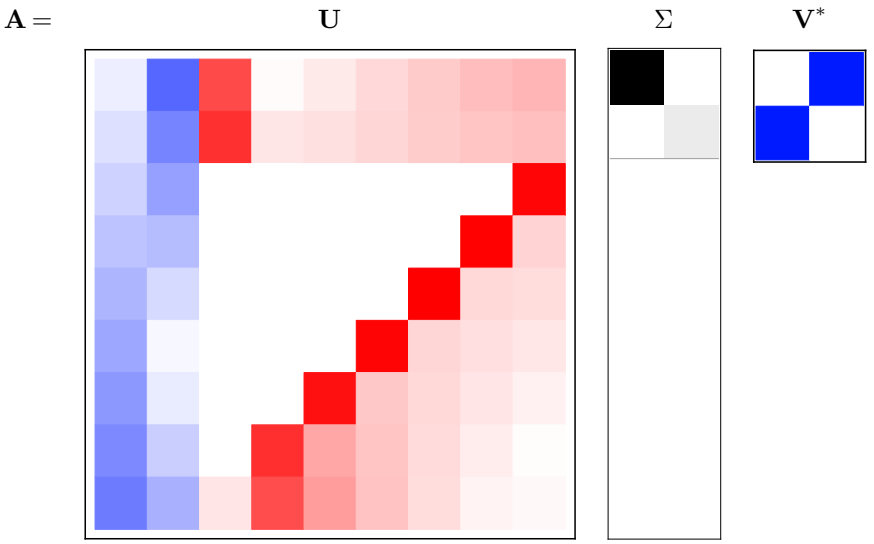


Table 4.2. *Singular value decomposition for the system matrix* $\boldsymbol{A}$ *in* (4.1) *showing range and null space components.*

Table 4.3. *A summary of the residual errors and their contributions to* $\|r\|_2$.

k	r_k	r_k^2
1	−1.4	1.9
2	6.1	37.6
3	−3.6	12.7
4	−1.4	1.8
5	−6.3	40.3
6	−0.3	0.1
7	6.5	41.9
8	9.7	93.7
9	−9.3	86.7

4.3 QR Decomposition

Resolve the range space $\mathcal{R}(\mathbf{A})$ into an orthonormal basis.

$$\mathbf{A} = \mathbf{QR}$$

$$a = \mathbf{R}^{-1}\mathbf{Q}^{*}T$$

4.3.1 Computing the QR Decomposition

$\mathbf{R} \in \mathbb{R}^{\rho \times \rho}$

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{bmatrix} = \begin{bmatrix} \nu_1 & q_1^2 a_2 \\ 0\nu_2 & \end{bmatrix}$$

The norm of the first column vector:

$$r_{11} = \|a_1\|_2 = 3 \tag{4.7}$$

Normalized column vector

$$q_1 = \frac{a_1}{r_{11}} = \frac{1}{3}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \tag{4.8}$$

Next scale factor in row 1:

$$r_{12} = q_1^* a_2 = 15 \tag{4.9}$$

$$q_2 = a_2 - r_{12} q_1 \tag{4.10}$$

$$\mathbf{Q} = \begin{bmatrix} & \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} & & \begin{bmatrix} -4 \\ -3 \\ -2 \\ -1 \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} \\ \frac{1}{3} & & \frac{1}{2\sqrt{15}} & \end{bmatrix}$$

$$\mathbf{A} = \mathbf{Q} \mathbf{R},$$

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \\ 1 & 7 \\ 1 & 8 \\ 1 & 9 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} & \frac{1}{2\sqrt{15}} \begin{bmatrix} -4 \\ -3 \\ -2 \\ -1 \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} \end{bmatrix} \begin{bmatrix} 3 & 15 \\ 0 & 2\sqrt{15} \end{bmatrix}$$

Chapter 5

Modal Example III Observations

The demonstration problem provides a platform to explore deeper aspects of the least squares fit.

5.1 Invariances

Invariance is a touchstone which allows for elegant discrimination of theoretical and numerical analysis. The critical insight is this: complicated problems have complicated solutions, but they retain simple invariances. Simple properties afford simple tools for checking complicated problems.

5.1.1 Translation Invariance

Intuitively, a translation of the coordinate system will not change the temperature measurements, which means the gradient of the temperature is invariant under translation of the coordinate system. For the example in figure 3.1, the temperature gradient was measured over the length of a 10 cm bar. For convenience, the end of the bar in contact with the ice bath was taken as the origin. But just as readily, that point could have been taken as $x = 5$ cm; the end of the bar touching the boiling water bath is then 15 cm. The physics is unchanged. For an arbitrary shift of x_* the gradient is invariant:

$$\nabla T = \frac{T_1 - T_0}{(x_1 + x_*) - (x_0 + x_*)} = \frac{T_1 - T_0}{x_1 - x_0}$$

How should the solution change? We insist the slope (gradient) remain unchanged. The origin shifts from $x_0 = 0$ to $x_0 = -x_*$, the intercept shift according to

$$y(-x_0) = a_0 - a_1 x,$$

that is, $a_0 \to a_0 - a_1 x_*$.

Translation

Formally evaluate how the least squares solution in (**??**) transforms under the translation

$$x_k \to x_k + x_*$$

The position vector x can be written as the sum of the original vector plus the translation:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \to \begin{bmatrix} x_1 + x_* \\ x_2 + x_* \\ \vdots \\ x_m + x_* \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} + \begin{bmatrix} x_* \\ x_* \\ \vdots \\ x_* \end{bmatrix};$$

in vector notation this is

$$x \to x + \mathbf{1}x_*. \tag{5.1}$$

Demonstrations

First compute the transformation rules on the inner products using (**??**):

$$\begin{aligned} &\mathbf{1}^{\mathrm{T}}\mathbf{1} \to \mathbf{1}^{\mathrm{T}}\mathbf{1}, \\ &\mathbf{1}^{\mathrm{T}}T \to \mathbf{1}^{\mathrm{T}}T, \\ &\mathbf{1}^{\mathrm{T}}x \to \mathbf{1}^{\mathrm{T}}x + x_* \left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right), \\ &x^{\mathrm{T}}x \to x^{\mathrm{T}}x + 2x_* \left(\mathbf{1}^{\mathrm{T}}x\right) + x_*^2 \left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right), \\ &x^{\mathrm{T}}T \to x^{\mathrm{T}}T + x_* \left(\mathbf{1}^{\mathrm{T}}T\right). \end{aligned} \tag{5.2}$$

Computation of the transformation of the intercept a_0 and slope a_1 requires computation of the invariance of the determinant: Invariance of determinant (4.2):

$$\begin{aligned} \Delta &= \left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)\left(x^{\mathrm{T}}x\right) - \left(\mathbf{1}^{\mathrm{T}}x\right)^2, \\ &\to \left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)\left(x^{\mathrm{T}}x + 2x_*\left(\mathbf{1}^{\mathrm{T}}x\right) + x_*^2\left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)\right) - \left(\mathbf{1}^{\mathrm{T}}x + x_*\left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)\right)^2, \\ &= \Delta + \left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)\left(2x_*\left(\mathbf{1}^{\mathrm{T}}x\right) + x_*^2\left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)\right) - \left(2x_*\left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)\left(\mathbf{1}^{\mathrm{T}}x\right)\right) - \left(x_*\left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)\right)^2, \\ &= \Delta. \end{aligned}$$

Thus the determinant is invariant under translation.

The solution parameters transform as Solutions in (4.3).

$$\begin{aligned} a_1 &= \Delta^{-1}\left(\left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)\left(x^{\mathrm{T}}T\right) - \left(\mathbf{1}^{\mathrm{T}}x\right)\left(\mathbf{1}^{\mathrm{T}}T\right)\right), \\ &\to \Delta^{-1}\left(\left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)\left(x^{\mathrm{T}}T + x_*\left(\mathbf{1}^{\mathrm{T}}T\right)\right) - \left(\mathbf{1}^{\mathrm{T}}x + x_*\left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)\right)\left(\mathbf{1}^{\mathrm{T}}T\right)\right), \\ &= a_1 + \Delta^{-1}\left(x_*\left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)\left(\mathbf{1}^{\mathrm{T}}T\right) - x_*\left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)\left(\mathbf{1}^{\mathrm{T}}T\right)\right), \\ &= a_1 \end{aligned}$$

Computations

The slope is invariant, the intercept:

$$\begin{aligned} a_1 &\to a_1, \\ a_0 &\to a_0 - x_* a_1. \end{aligned}$$

Predictions for $x_* = 1$.

$$\begin{aligned} a_0 &\to -\frac{827}{180}, \\ a_1 &\to \frac{1129}{120}. \end{aligned} \tag{5.3}$$

$$\begin{aligned} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} &= \Delta^{-1} \begin{bmatrix} \sum x_k^2 & -\sum x_k \\ -\sum x_k & \sum 1 \end{bmatrix} \begin{bmatrix} \sum T_k \\ \sum T_k x_k \end{bmatrix}, \\ &= 540^{-1} \begin{bmatrix} 384 & -54 \\ -54 & 9 \end{bmatrix} \frac{1}{10} \begin{bmatrix} 4667 \\ 33\,647 \end{bmatrix}, \\ &= \frac{1}{360} \begin{bmatrix} 1733 \\ 3387 \end{bmatrix}. \end{aligned}$$

In agreement with (**??**).

Visuals

Compare to figure 3.2

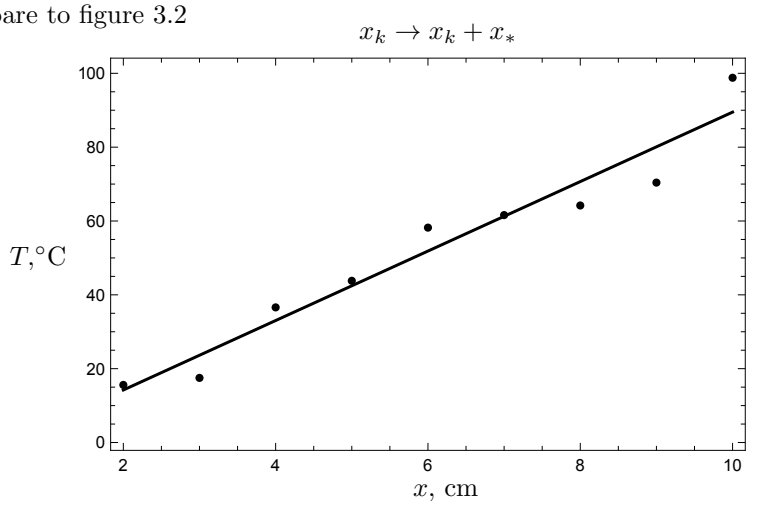


Figure 5.1. *Solution after translation along x axis: the slope is invariant.*

Compare to 3.3
Compare to figure 3.4.1

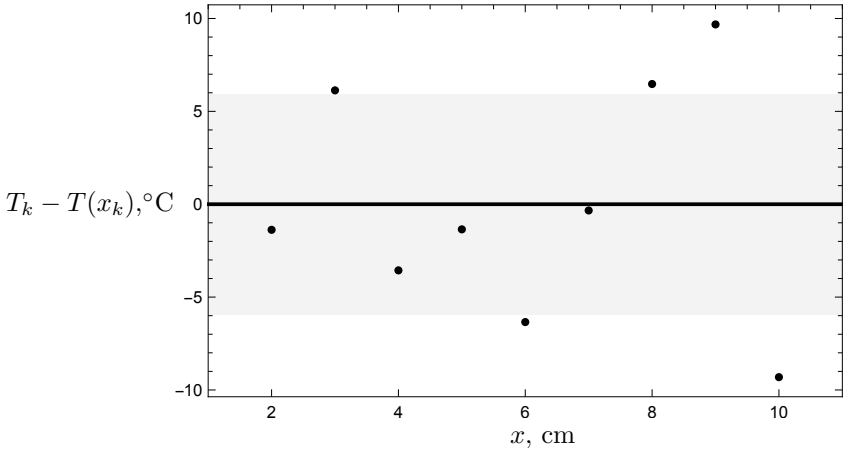


Figure 5.2. *Scatter plot of residual errors after translation along x axis: the residuals are invariant.*

5.1.2 Reflection Invariance

A demonstration of the reflection method of §2.5.

$$\sum_{k=1}^{m} r^2 = \sum_{k=1}^{m} (-r)^2$$

table 3.3

5.2 Fitting To Higher Orders

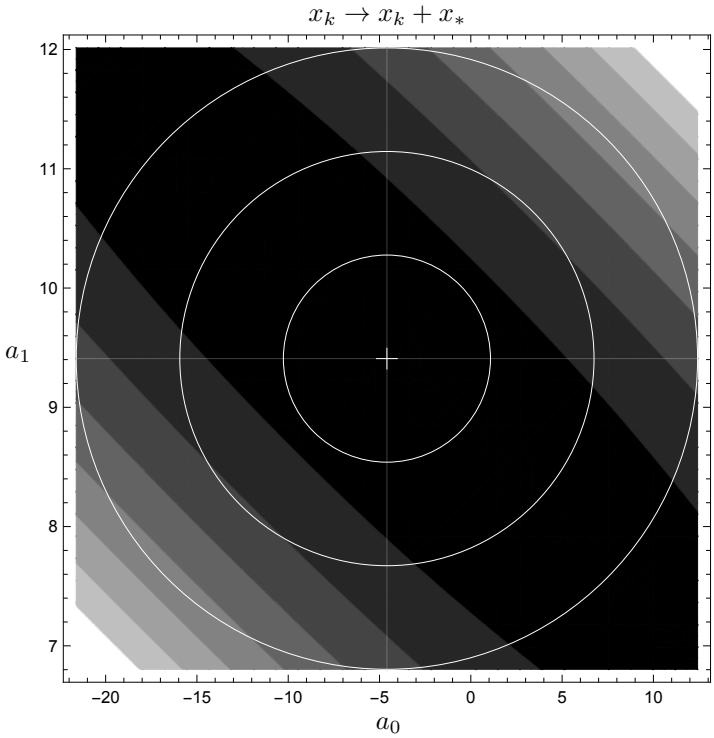


Figure 5.3. *Contour plot of merit function after translation along x axis: a_1 is invariant.*

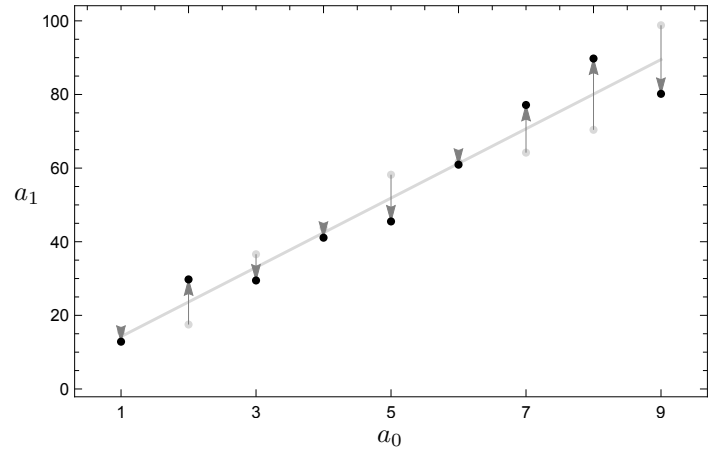


Figure 5.4. *The merit function for the learning curve showing the minimum and the value.*

Part II

Applications: Nonlinear Problems

Chapter 6

Linearization

6.1 Powers Laws and Exponentials

Exponential relationships, such as radioactive decay,

$$y = ae^{bx},$$

and power laws, like the learning curve,

$$y = ax^b \tag{6.1}$$

are inherently nonlinear. As such, they are outside of the scope of this work. But the problem is that the Internet has helped create and maintain the myth that such problems can be tamed and made linear. This is, empathically, not true.

So this chapter is about the proper way to handle power law and exponential relationships. Along the way, it will become clear why faux linearization does not work. In short, the answer is that the linearization *changes the problem*. That is, it produces a different problem with a different solution.

$$a_{LS} = \left\{ a \in \mathbb{R}^2 \colon \left\| y_k - ax^b \right\|_2^2 \text{ is minimized} \right\}$$

6.2 Learning Curve

The trial function

$$y(x) = ax^b, \quad x \in \mathbb{R}^+.$$

If we restrict $x \in \mathbb{R}^+$, then $b \in \mathbb{R}$, avoiding the indeterminate form 0^0. In practice $x \in \mathbb{N}$. The variable a is an initial value (at $x = 1$), and b is a rate parameter.

The merit function is

$$M(a, b) = \sum_{k=1}^{m} \left(y_k - ak^b \right)^2 \tag{6.2}$$

k	y_k	$y(k)$	r_k	k	y_k	$y(k)$	r_k
1	419.115	371.9	47.2149	27	119.695	128.685	–8.98984
2	251.263	297.505	–46.2419	28	131.931	127.187	4.74458
3	291.329	261.092	30.2371	29	126.702	125.758	0.944397
4	208.272	237.992	–29.7202	30	107.923	124.392	–16.4689
5	198.577	221.492	–22.9146	31	121.735	123.086	–1.35074
6	172.583	208.863	–36.2804	32	136.803	121.834	14.969
7	202.107	198.749	3.35859	33	97.0772	120.633	–23.5554
8	169.911	190.384	–20.4736	34	110.727	119.479	–8.75175
9	175.674	183.299	–7.62479	35	132.09	118.369	13.7215
10	191.393	177.185	14.2084	36	94.4017	117.3	–22.8979
11	152.023	171.83	–19.8067	37	111.213	116.269	–5.05649
12	183.7	167.082	16.6184	38	113.38	115.275	–1.89502
13	157.806	162.831	–5.02491	39	112.433	114.315	–1.88186
14	142.933	158.991	–16.0584	40	123.706	113.387	10.319
15	185.178	155.498	29.6797	41	125.563	112.489	13.0742
16	172.109	152.3	19.8089	42	101.05	111.619	–10.5695
17	174.763	149.356	25.4069	43	101.805	110.777	–8.9723
18	151.21	146.632	4.57821	44	107.492	109.96	–2.46808
19	132.162	144.101	–11.9391	45	89.7489	109.167	–19.4182
20	125.188	141.741	–16.5524	46	107.598	108.397	–0.799046
21	144.023	139.531	4.49148	47	117.91	107.649	10.261
22	117.048	137.457	–20.4084	48	94.22	106.922	–12.7018
23	125.011	135.503	–10.4919	49	108.116	106.214	1.90216
24	144.994	133.659	11.3349	50	108.17	105.526	2.64485
25	126.14	131.914	–5.77349	51	116.083	104.855	11.2277
26	146.929	130.258	16.6713	52	123.93	104.201	19.7289

To advance the discussion, consider the sample set of data for 52 points in table **??**. What does the surface of the merit function in (**??**) look like in solution space? Very much like all the other merit functions so far. There is a unique minimum, a distinct feature of the landscape. The question becomes how to find it.

6.2.1 Problem Statement

The problem stated in table **??** embodies the major points discussed up to this point and displays a few details that will become clear as the discussion advances.

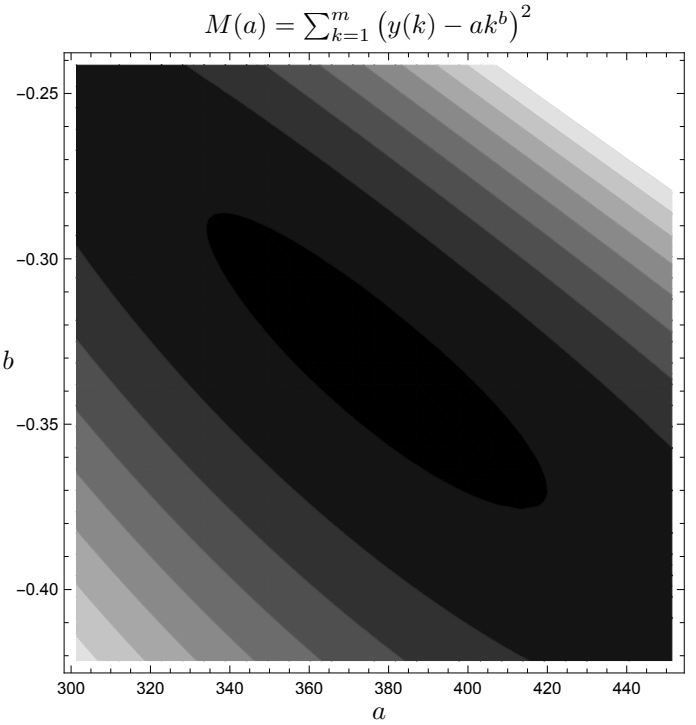


Figure 6.1. *Merit function for the learning curve in solution space.*

6.2.2 Solution

Without a linear system, attack with calculus as in §3.2.2. Find the roots of the gradient $\nabla M(a, b) = 0$:

$$\partial_a M = -2 \sum_{k=1}^{m} \left(y_k - ak^b\right) k^b = 0,$$
$$\partial_b M = -2 \sum_{k=1}^{m} \left(y_k - ak^b\right) ak^b \ln k = 0.$$

Look at the equations in explicit form in table **??**. There is no linear system which isolates the solution parameters in a fashion such as this:

$$\mathbf{A} \begin{bmatrix} a \\ b \end{bmatrix} = \zeta.$$

The condition $\partial_a M = 0$ allows the separation of the solution parameters.

$$a(b) = \frac{\sum y_k k^b}{\sum k^{2b}}$$

This permits removal of one variable, and reduces this merit function to a single

Table 6.1. *Problem statement for learning curve.*

trial function	$y(x) = ae^{b \ln x}$	$a \in \mathbb{R}^+$, $b \in \mathbb{R}$
residual error	$r_k = y(k) - ak^b$	
merit function	$M(a,b) = \sum_{k=1}^{m} \left(y(x) - ak^b\right)^2$	
unit numbers	k, $k = 1\!:\!m$	
measurements	y_k, $k = 1\!:\!m$	
results	a	initial value
	b	learning rate
# of measurements	$m = 52$	
# of parameters	$n = 2$	
system matrix	no system matrix – problem is nonlinear	
linear system	no linear system – problem is nonlinear	
unperturbed solution	$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 371.9 \\ -0.322 \end{bmatrix}$	80% rate
input data	table **??**	

Table 6.2. *The simultaneous conditions defining* $\nabla M(a,b) = 0$.

$\partial_a M$			$\partial_b M$		
$a1^{2b}$	$=$	$y_1 1^b$	$a^2 1^{2b}$	$=$	$ay_1 1^b \ln 1$
$a2^{2b}$	$=$	$y_2 2^b$	$a^2 2^{2b}$	$=$	$ay_2 2^b \ln 2$
	$\vdots$			$\vdots$	
$a52^{2b}$	$=$	$y_{52} 52^b$	$a^{52} 52^{2b}$	$=$	$ay_{52} 52^b \ln 52$

parameter, b:

$$\hat{M}(b) = \sum \left(y(x) - \frac{\sum y_k k^b}{\sum k^{2b}} k^b \right)^2 \tag{6.3}$$

where summations from $k = 1$ to 52 are implied.

Figure **??** shows the plot of $\hat{M}(b)$ with the minimum marked. The nonlinear least squares problem in two dimensions is now reduced to a minimization problem in one dimension. The solutions are posted in table **??**. The curve in figure **??** represents a slice through the surface of the merit function in figure **??**. Such a slice is shown in figure **??**.

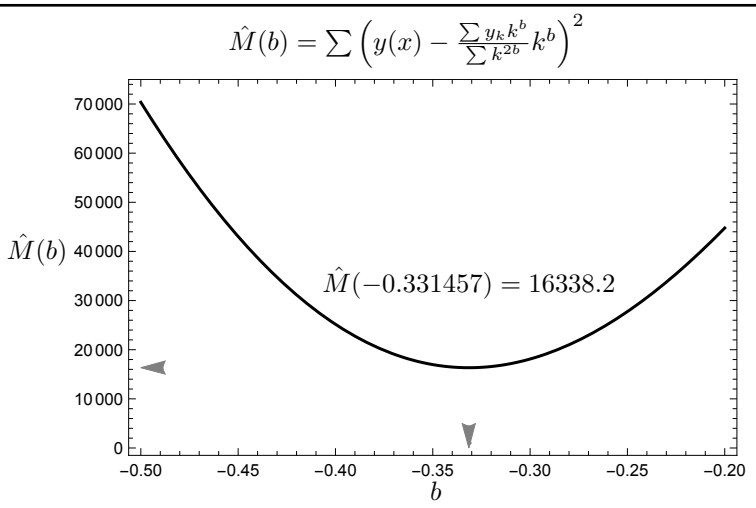


Figure 6.2. *Merit function constrained to one parameter b.*

6.2.3 Results

The minimizer is quoted to full double precision for debugging purposes:

$$b_{LS} = -0.33145698171782384,$$

the initial value is

$$a(b_{LS}) = 376.3700704535057,$$

and the minimum of the merit function is

$$M\left(a_{LS}, b_{LS}\right) = \hat{M}\left(b_{LS}\right) = 16338.235321721559.$$

The results and residuals are posted in figure **??**.

The merit function in figure 2.5 has a new feature. In addition to the least squares solution and the error ellipses, the unperturbed solution is plotted with and "×". This is the ideal solution used to generate the raw data before the noise is added. The least squares solution will converge to this point as the magnitude of the error decreases.

6.3 What Not To Do

Folklore promotes the idea of transforming the problem to a more amenable form. And certainly transformation is a powerful and useful mathematical technique. But the popular transformation complicates the problem and provides an incorrect answer.

6.3.1 Logarithmic Transform

While the logarithmic transformation is exact,

$$y = ax^b \qquad \Rightarrow \qquad \ln y = \ln a + b \ln x,$$

$$M(a,b) = \sum_{k=1}^{m} \left(y(x) - ak^b\right)^2$$

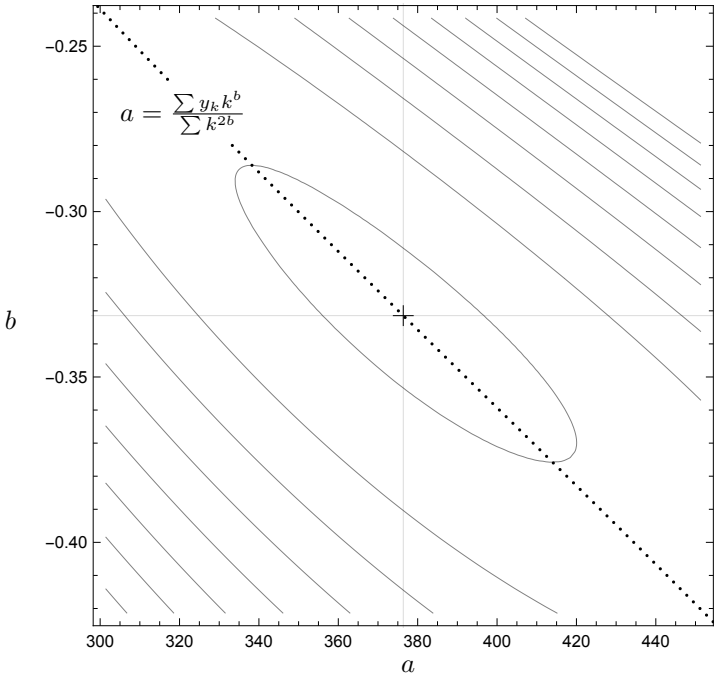


Figure 6.3. *Merit function for the learning curve in solution space showing the constrained a parameter as a dotted line.*

$$y(x) = ax^b$$

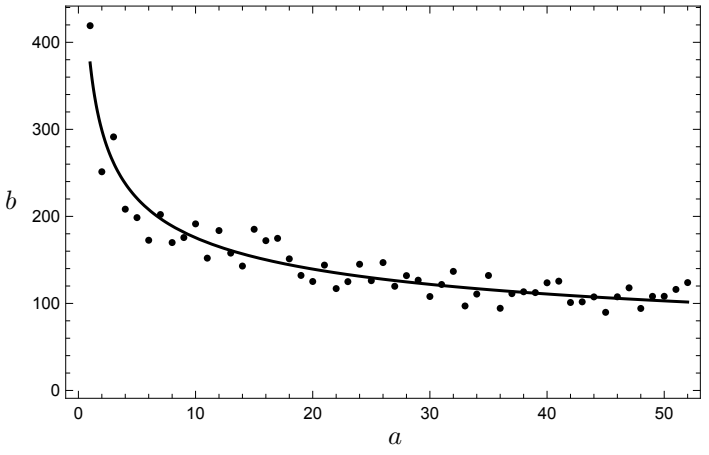


Figure 6.4. *Solution for equations* (**??**) *and* (**??**) *using data in table* **??**.

the application is inappropriate because the problem is not exact. If there is no noise in the data set, the above transformation will preserve the solution. However, if the problem is exact, there is no need to employ least squares. One could take

Table 6.3. *Results for learning curve analysis.*

fit parameters	a	initial value
	b	rate
solution function	$y(x) = ax^b$	$[L]$
solution error		
computed solution	$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 376.37 \\ -0.331457 \end{bmatrix} \pm \begin{bmatrix} ? \\ ? \end{bmatrix}$	
unperturbed solution	$\begin{bmatrix} \tilde{a} \\ \tilde{b} \end{bmatrix} = \begin{bmatrix} 376 \\ -0.322 \end{bmatrix}$	
r^*r	16 338	
problem statement	table **??**	
input data	table **??**	
plots	figure **??**	1. data and solution
	figure **??**	2. residual errors
	figure 2.5	3. merit function

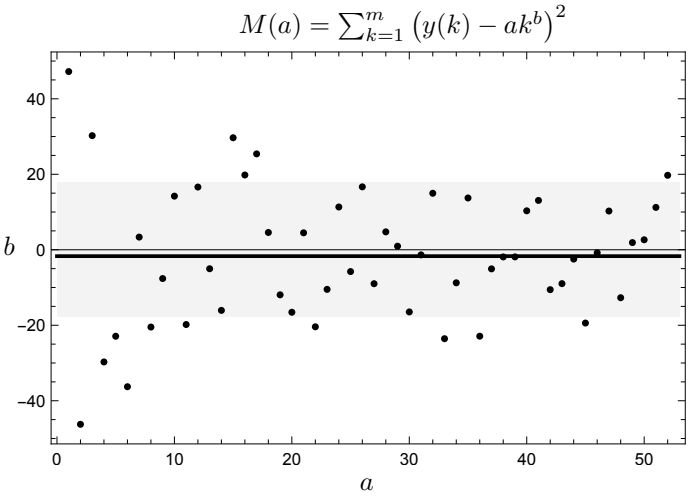


Figure 6.5. *The residual errors in figure* **??**.

the ratio of two distinct data points,

$$\frac{y_j}{y_k} = \left(\frac{j}{k}\right)^b ,$$

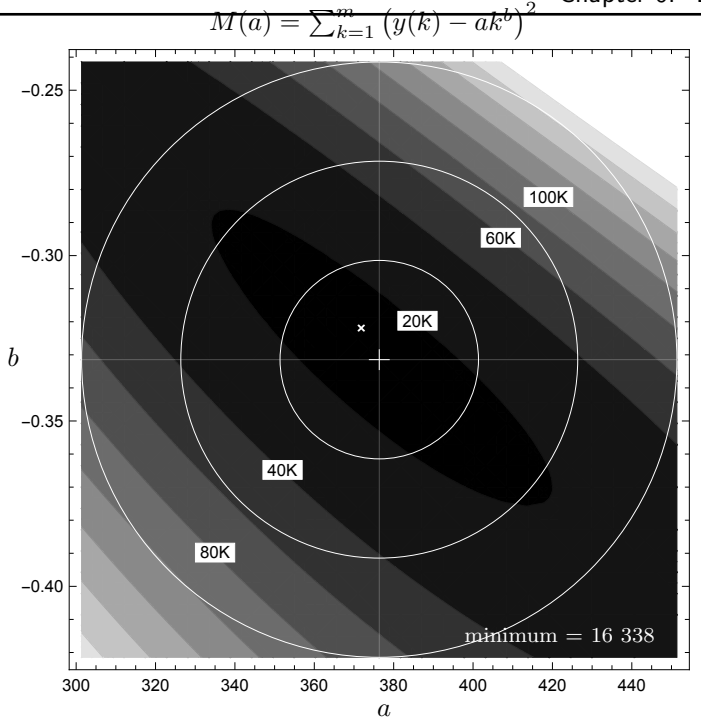


Figure 6.6. *Minimization of the merit function for the learning curve.*

and compute

$$b = \frac{\ln\left(y_j/y_k\right)}{\ln\left(j/k\right)}.$$

Again, there is no need for least squares methods.

When the data contains noise, and the choice is to use least squares, the logarithmic transformation corrupts the data. Look at the transformation of trial function in the presence of noise, that is, with nonzero residual error:

$$y_k = ak^b + r_k \qquad \Rightarrow \qquad \ln y_k = \ln\left(ak^b + r_k\right).$$

The logarithmic transform is not linear and does not separate the parameters a and b in the presence of error.

Because this malady is so common, the solution invites belaboring. The function $\ln y = \ln a + b \ln x$ is linear in the new coordinates $\ln y$ and $\ln x$ but only when there is no noise – only when there is no need for least squares arbitration. The presence of noise creates an additive term and the logarithmic transformation creates a different problem which has a different solution.

6.3.2 Linear Transformation

A transformation T is linear if and only if

$$T(x + \alpha y) = T(x) + \alpha T(y).$$

The logarithmic transformation fails this primal test:

$$\ln(x + \alpha y) \neq \ln(x) + \alpha \ln(y)$$

The faux solution is

$$\begin{aligned} \ln a &= 5.88094 \qquad \rightarrow \qquad a = 358.146, \\ b &= -0.314929. \end{aligned} \tag{6.4}$$

Note that the value of the merit function is higher than the true minimum:

$$M(358.146, -0.314929) = 17005 > 16338. \tag{6.5}$$

The eyeball is a poor instrument for distinguishing between the correct solution in table 3.3 and (**??**). However, the merit function sharply delineates the two.

The merit shows the solution (**??**) marked with ⊞, distinctly away from the minimum of the merit function as seen in equation (**??**) and figure 2.5.

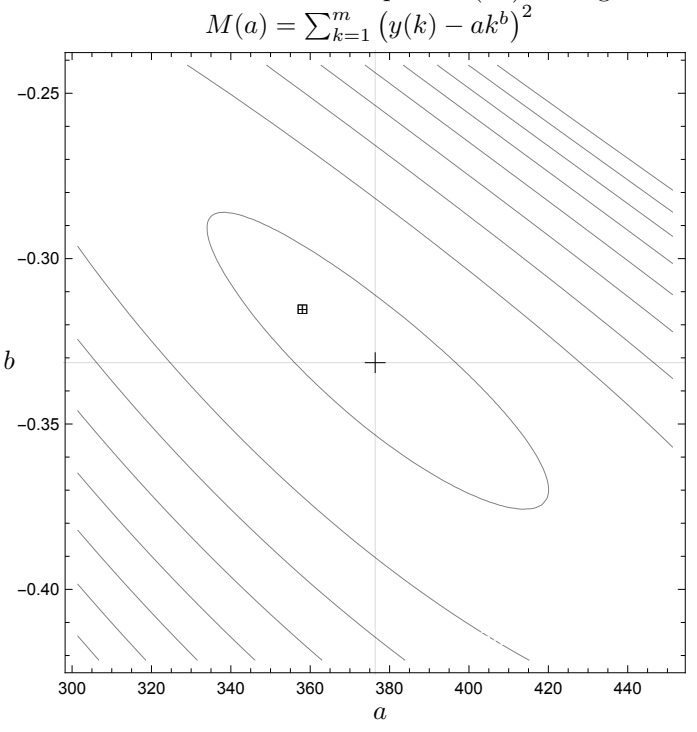


Figure 6.7. *The merit function for the learning curve showing the minimum and the value .*

6.3.3 Reflection Test Fails

6.4 Radioactive Decay

6.4.1 Theory

6.4.2 Problem Statement

Table 6.4. *Problem statement for radioactive decay.*

trial function	$c(t) = ae^{bt}$	counts
residual error	$r_k = c_k - ae^{bt_k}$	counts
merit function	$M(a) = \sum_{k=1}^{m} \left(c_k - ae^{bt_k}\right)^2$	counts^2
measurements	t_k, $k = 1{:}m$	time, s
	c_k, $k = 1{:}m$	counts
results	$a \pm \epsilon_a$	initial value
	$b \pm \epsilon_b$	power
# of measurements	$m = 10$	rows in $\mathbf{A}$
# of parameters	$n = 2$	columns in $\mathbf{A}$
system matrix	$\mathbf{A} \in \mathbb{R}_2^{10\times 2}$	
linear system	$\begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} T_1 \\ \vdots \\ T_m \end{bmatrix}$	
ideal solution	$\begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 10 \end{bmatrix}$	
input data	table 3.2	

$$C = C_0 e^{-t/\tau}$$

6.4.3 Results

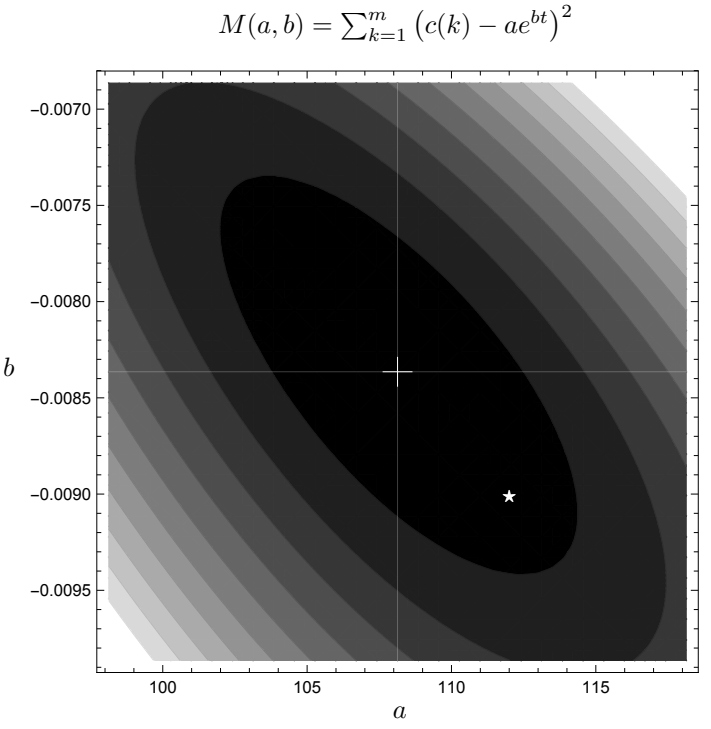
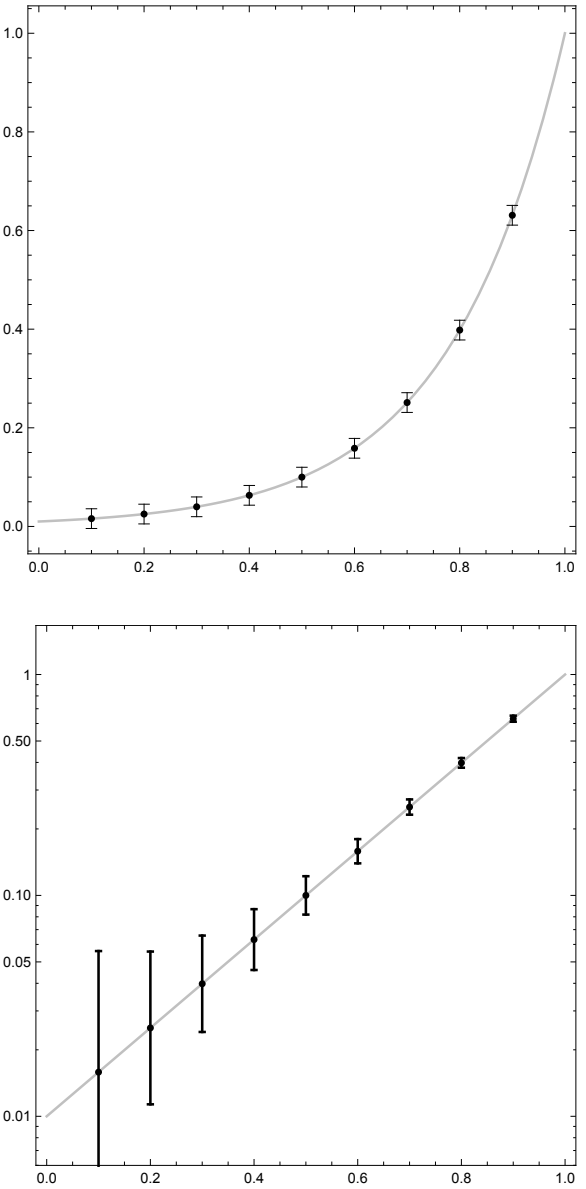


Figure 6.8. *The merit function for the learning curve showing the minimum and the value .*

Table 6.5. *Logarithmic scaling distorts errors.*



k	t	c
1	0	106
2	15	80
3	30	98
4	45	75
5	60	74
6	75	73
7	90	49
8	105	38
9	120	37
10	135	22

Table 6.6. *Results for radioactive decay.*

fit parameters	a_0	intercept, °C
	a_1	slope, °C / cm
solution function	$T(x) = a_0 + a_1 x$	°C
solution error	$\epsilon_T^2(x) = \epsilon_0^2 + x^2\epsilon_1^2 + a_1^2\epsilon_x^2$	°C
computed solution	$\begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 4.8 \\ 9.41 \end{bmatrix} \pm \begin{bmatrix} 4.9 \\ 0.87 \end{bmatrix}$	
ideal solution	$\begin{bmatrix} \tilde{a}_0 \\ \tilde{a}_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 10 \end{bmatrix}$	
r^*r	316.6	
curvature matrix $(\mathbf{A}^*\mathbf{A})^{-1}$	$\frac{1}{180}\begin{bmatrix} 95 & -15 \\ -15 & 3 \end{bmatrix}$	
problem statement	table 3.1	
input data	table 3.2	
plots	figure 3.2	1. data and solution
	figure 3.3	2. residual errors
	figure **??**	3. merit function

Chapter 7

Population Growth

In this section we take a nonlinear model for population growth and separate the linear and nonlinear terms.

7.1 Model

$$y(\tau) = a_1 + a_2\tau + a_3 e^{d\tau} \tag{7.1}$$

$$\mathbf{A}(d+\gamma) \neq \mathbf{A}(d) + \mathbf{A}(\gamma)$$

$$\begin{matrix} & \mathbf{A}(d) & & a & = & y \end{matrix}$$

$$\begin{bmatrix} 1 & \tau_1 & e^{d\tau_1} \\ 1 & \tau_2 & e^{d\tau_2} \\ 1 & \tau_3 & e^{d\tau_3} \\ 1 & \tau_4 & e^{d\tau_4} \\ 1 & \tau_5 & e^{d\tau_5} \\ 1 & \tau_6 & e^{d\tau_6} \\ 1 & \tau_7 & e^{d\tau_7} \\ 1 & \tau_8 & e^{d\tau_8} \end{bmatrix} \quad \begin{bmatrix} a_0 \\ a_1 \\ a_3 \end{bmatrix} \quad = \quad \begin{bmatrix} y_0 \\ y_1 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \end{bmatrix}$$

$$\min_{\substack{a\in\mathbb{R}^3 \\ d\in\mathbb{R}}} \left\| \mathbf{A}(d) \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} - y \right\|_2^2 \tag{7.2}$$

Table 7.1. *Problem statement for population model with linear and exponential growth.*

trial function	$y(\tau) = a_0 + a_1\tau + a_2 e^{d\tau}$	$a \in \mathbb{R}^3$ $d \in \mathbb{R}$
merit function	$M(a,d) = \sum_{k=1}^{m} \left(y_k - a_1 + a_2\tau + a_3 e^{d\tau_k}\right)^2$	
# measurements	$m = 8$	
# parameters	$n = 4$	
rank defect	$\rho = n$	overdetermined
input data	(τ_k, y_k), $k = 1:8$	table **??**
results	a_0	constant
	a_1	linear
	a_2	exponential
	d	power term
residual error	$r = \mathbf{A}^{\dagger} b - \Delta$	
linear system	$\mathbf{A}(d) \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = y$	

7.2 Problem Statement

$$r^2 = \left\| \mathbf{A}(d) \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} - y \right\|_2^2 \tag{7.3}$$

7.3 Data

7.4 Example

year $= 1900 + 10(\tau - 1)$

7.5 Polynomials

There is the model we choose and the model which nature chooses. Are they the same?

Table 7.2. *Data v. prediction.*

year	census	fit	r	rel. error
1900	76.00	77.51	1.51	2.0%
1910	91.97	90.98	−0.99	−1.1%
1920	105.71	104.87	−0.84	−0.8%
1930	122.78	119.48	−3.29	−2.7%
1940	131.67	135.36	3.69	2.8%
1950	150.70	153.46	2.76	1.8%
1960	179.32	175.45	−3.87	−2.2%
1970	203.24	204.26	1.029	0.5%

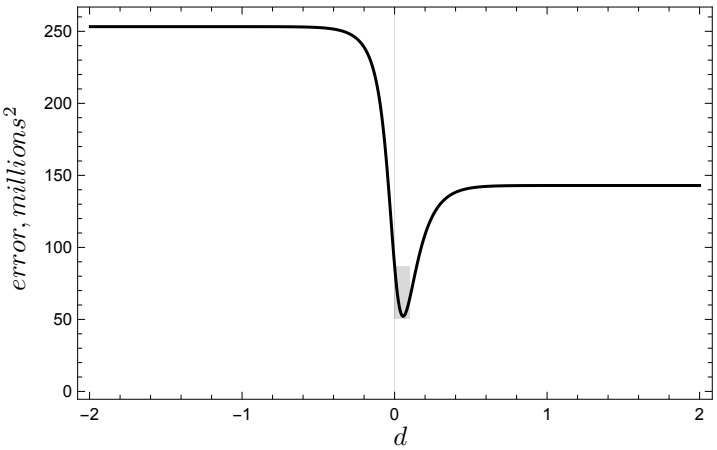


Figure 7.1. *Residual error for* (**??**) *with* a_1, a_2, *and* a_3 *at optimal values.*

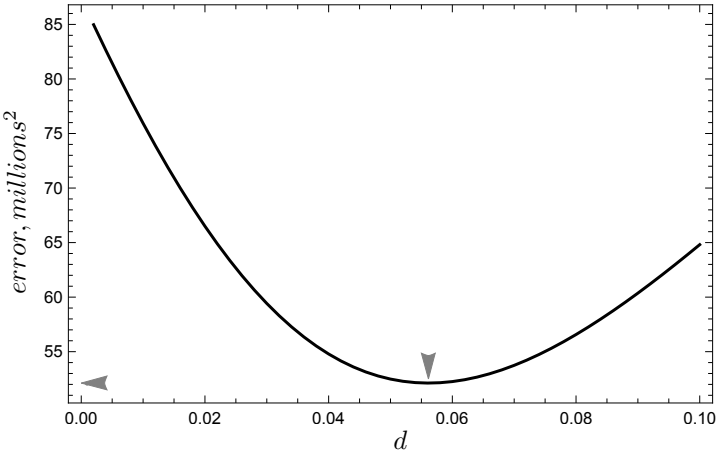


Figure 7.2. *Residual error for a_1 and a_2 fixed at optimal values.*

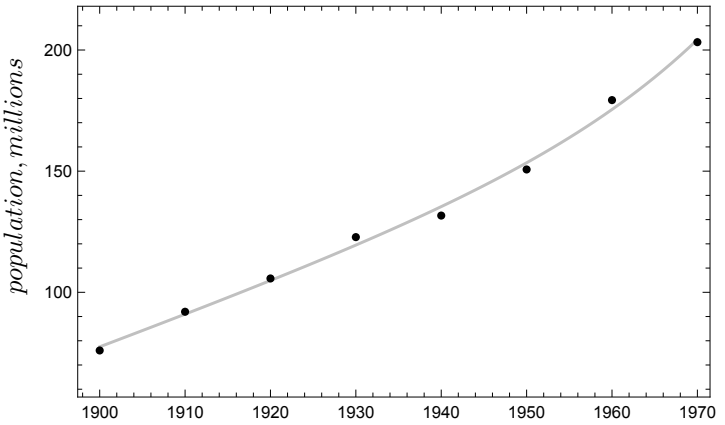


Figure 7.3. *Solution plotted against data.*

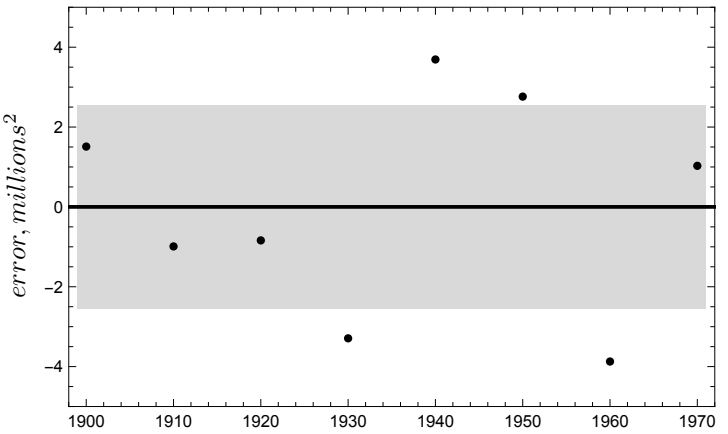


Figure 7.4. *Scatterplot of residual errors.*

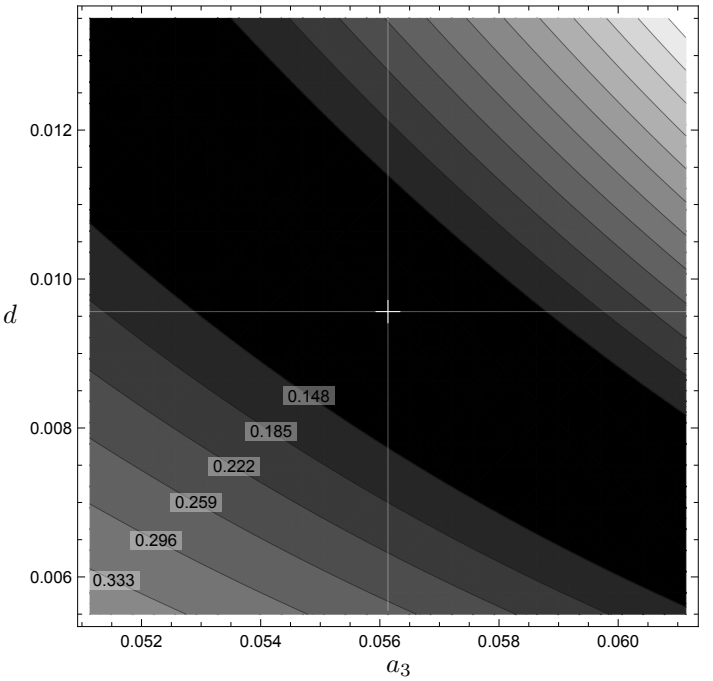


Figure 7.5. *The merit function with a_1 and a_2 fixed at best values showing least squares solution (center cross).*

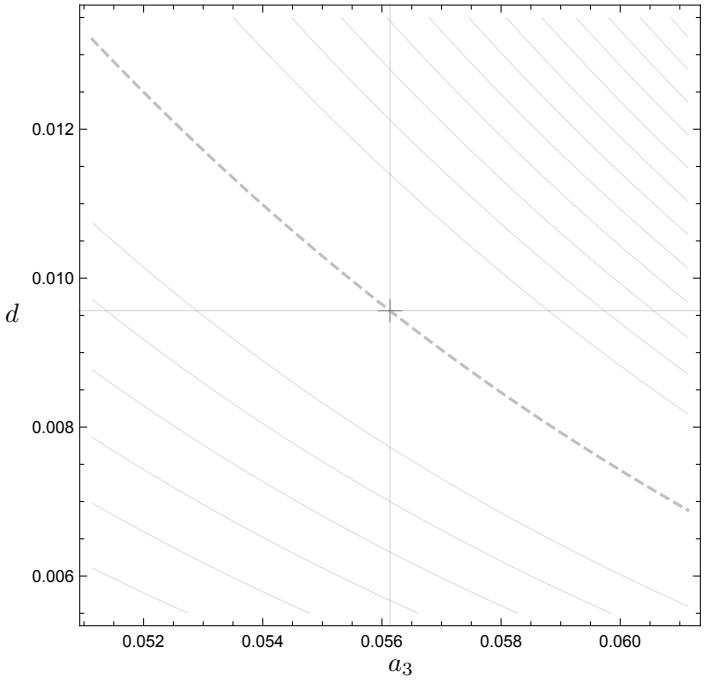


Figure 7.6. *The merit function with a_1 and a_2 fixed at best values showing least squares solution (center cross) and the (dashed line).*

Table 7.3. *Results for census analysis*

fit parameters	$c = \begin{bmatrix} 0.010 \\ 0.0170 \\ 0.0096 \end{bmatrix} \pm \begin{bmatrix} 0.031 \\ 0.0014 \\ 0.0020 \end{bmatrix}$
	$d = 0.056136 \pm ?.?$
$r^* r$	55.12
$\sum r_k$	-6.2×10^{-14}
σ_r	2.55
a	$\begin{bmatrix} 0.5397 & -0.0188 & 0.0165 \\ -0.0188 & 0.0011 & -0.0014 \\ 0.0165 & -0.0014 & 0.0022 \end{bmatrix}$
plots	data vs fit: figure **??**
	residuals: figure **??**
	merit function in $\mathcal{R}(\mathbf{A}^*)$: figure **??**

Table 7.4. *Fitting the census data with low order polynomials.*

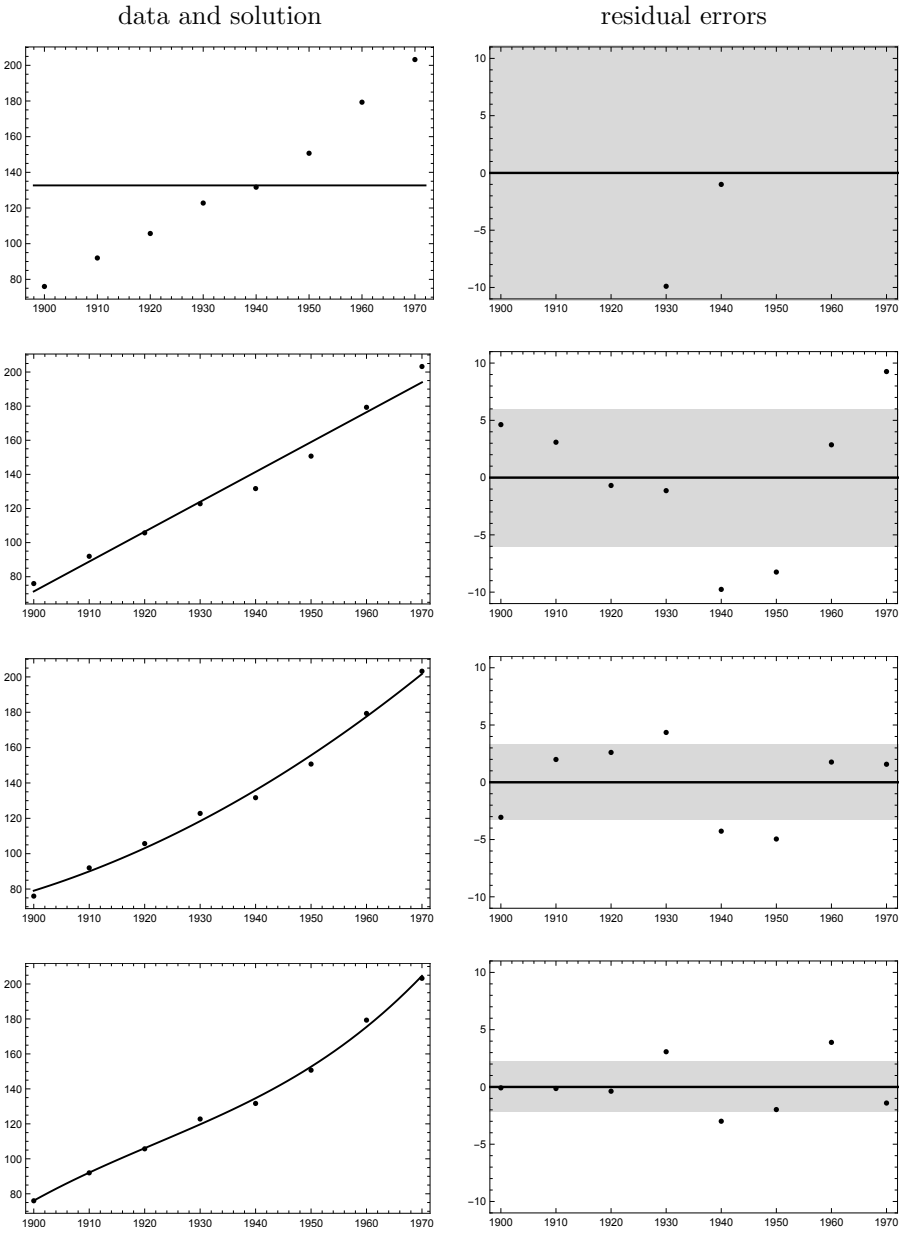
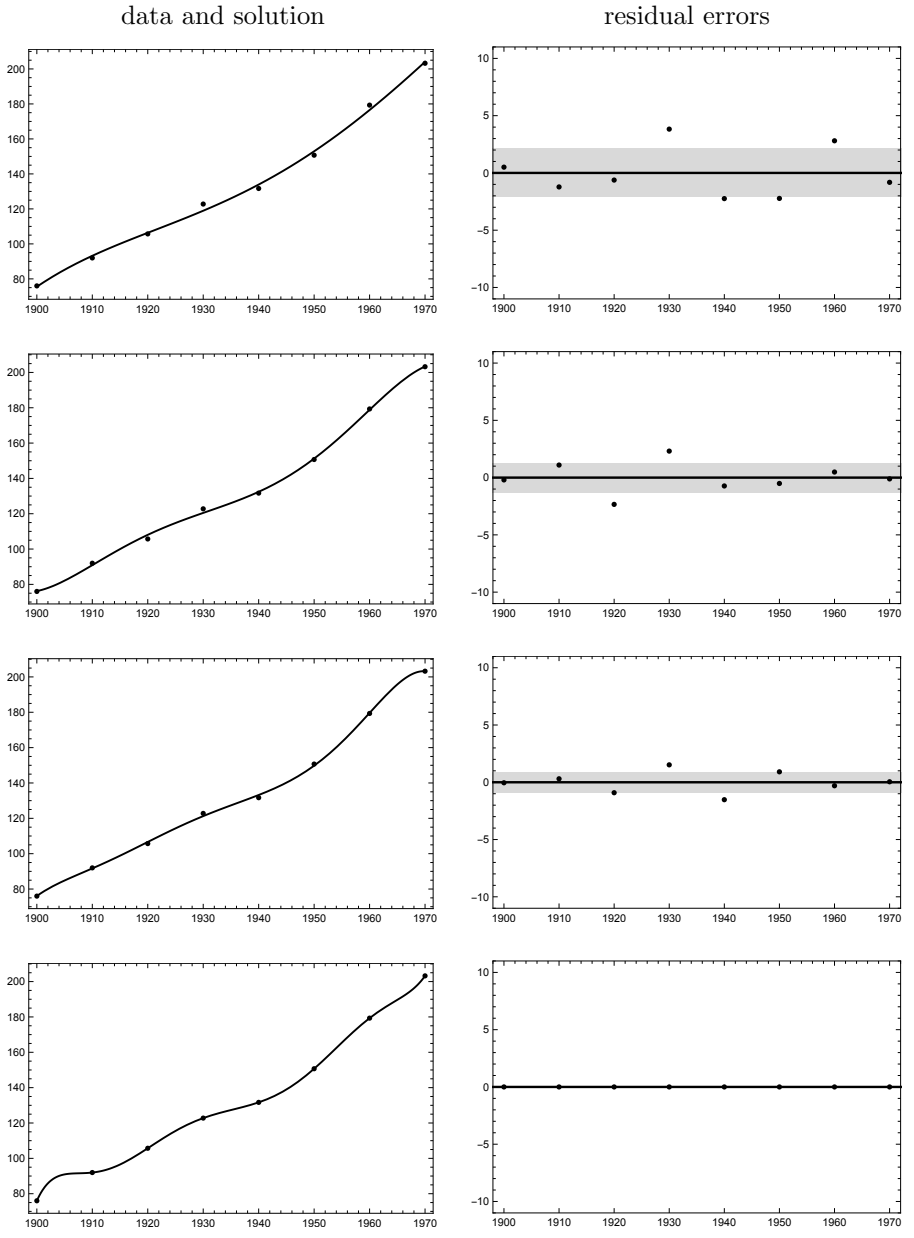


Table 7.5. *Fitting the census data with higher order polynomials.*



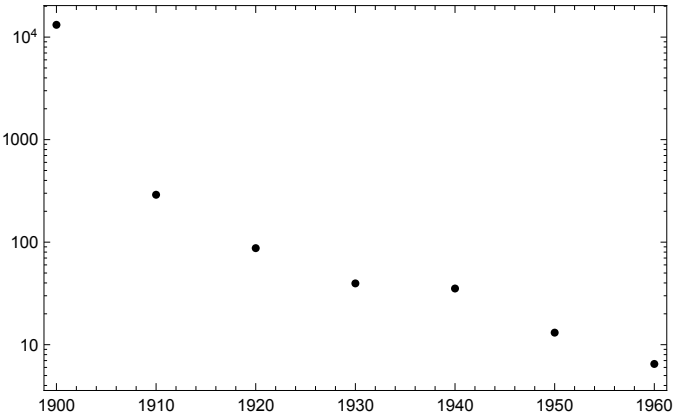


Figure 7.7. *Total error r*r by order of fit.*

Table 7.6. *Projections, by order of fit, for population in 2010.*

order	population (millions)
0	133
1	264
2	320
3	420
4	300
5	–928
6	–4,540
7	22,183

Part III

Backmatter

Bibliography

[1] Richard Bellman, *Introduction to matrix analysis*, SIAM, Society for Industrial and Applied Mathematics; 2^{nd} edition (1997).

[2] Philip R. Bevington, *Data Reduction and Error Analysis in the Physical Sciences*, McGraw-Hill (1969).

[3] Raymond H. Chan, and Chen Greif, and Diane P. O'Leary, *Milestones in matrix computation: Selected works of Gene H. Golub, with commentaries*, Oxford University Press (2007).

[4] James W. Demmel, *Applied numerical linear algebra*, SIAM, Society for Industrial and Applied Mathematics (1997).

[5] Gene H. Golub, and Charles Van Loan, *Matrix Computations*, 3^{rd} Edition. Johns Hopkins University Press (1996).

[6] Nicholas J. Higham, *Functions of Matrices: Theory and Computation*, SIAM, Society for Industrial and Applied Mathematics (2008).

[7] Roger A. Horn, and Charles R. Johnson, *Matrix analysis*, Cambridge University Press (1990).

[8] Roger A. Horn, and Charles R. Johnson, *Topics in Matrix analysis*, 3^{rd} Edition. Cambridge University Press (1991).

[9] Idris C. Mercer *Finding nonobvious nilpotent matrices*, (2005) `http://www.idmercer.com/nilpotent.pdf`

[10] Alan J. Laub, *Matrix analysis for scientists and engineers*, SIAM, Society for Industrial and Applied Mathematics (2005).

[11] Carl D. Meyer, *Matrix analysis and applied linear algebra*, SIAM, Society for Industrial and Applied Mathematics (2000).

[12] Gilbert Strang, *Linear Algebra and Its Applications*, SIAM, Society for Industrial and Applied Mathematics (2005).

[13] Lloyd N. Trefethen, and David Bau, *Numerical linear algebra*, SIAM, Society for Industrial and Applied Mathematics (2000).

[14] Eric W. Weisstein, "Characteristic Polynomial", from MathWorld–A Wolfram Web Resource.
http://mathworld.wolfram.com/CharacteristicPolynomial.html