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 solution 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 vectors of dimension n to 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 computation of approximations, the goal is to minimize the residual error and this work explores the minimal solutions under the 2−norm, the familiar norm of Pythagoras:

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

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, a solution with $\|\mathbf{A}x - b\| = 0$. 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

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

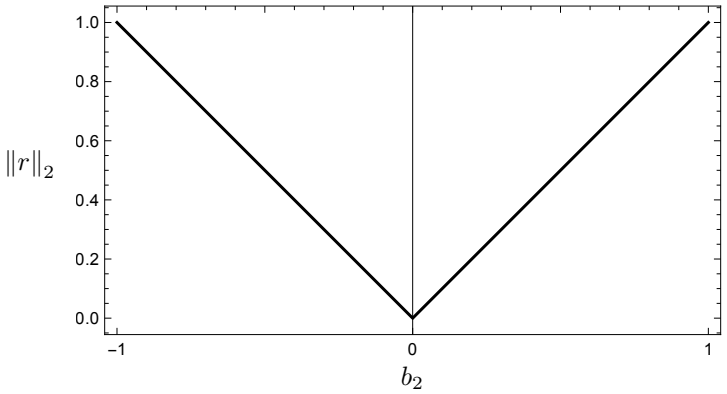


Figure 1.1. *The residual error* $\|r\|_2$ *given in* (1.5).

plotted in figure 1.1. This is the least possible error for the problem and (1.4) is the best solution. 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

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

The equality is attained when the data vector lies within the column space of $\mathbf{A}$.

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 ϕ.

Zonal Problem

In practice one measures the vector field and solves the inverse problem by computing values for the potential at the zone boundaries. The input and outputs are represented 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 $\phi(2)$.

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

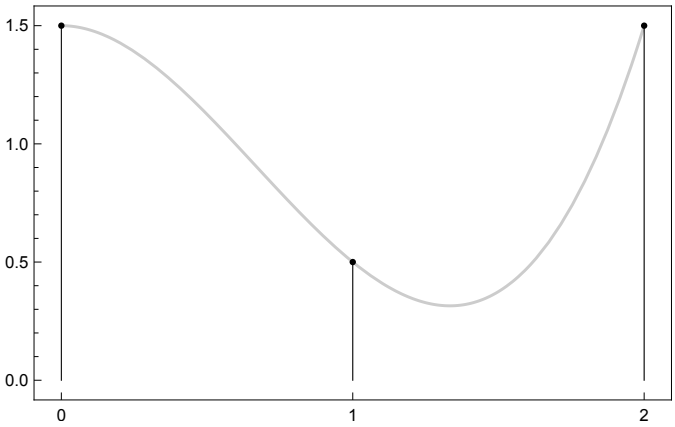


Figure 1.2. *Scalar function $\phi(x)$ (curve) and approximation φ_k (sticks).*

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 range space vectors, red null space vectors. In this way, the fundamental spaces spring to life.

There is a continuum of solutions due to the fact that

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

An quick demonstration confirms the red vector is in the null space:

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

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:

$$y(x)=a_0+a_1x$$

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

$$\begin{aligned}a_0+a_1x_1&=y_1\\ &\vdots\\ a_0+a_1x_m&=y_m.\end{aligned}$$

The goal is to simultaneously solve the set of equations.

Modal Solution

The first step is to compose the system

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

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:

$$\begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix} = \left((\mathbf{1}^{\mathrm{T}}\mathbf{1})(x^{\mathrm{T}}x) - (\mathbf{1}^{\mathrm{T}}x)^2 \right)^{-1} \begin{bmatrix} x^{\mathrm{T}}x & -\mathbf{1}^{\mathrm{T}}x \\ -\mathbf{1}^{\mathrm{T}}x & \mathbf{1}^{\mathrm{T}}\mathbf{1} \end{bmatrix} \begin{bmatrix} \mathbf{1}^{\mathrm{T}}y \\ x^{\mathrm{T}}y \end{bmatrix}. \tag{1.8}$$

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 General Least Squares Problem and Solution

Emboldened by solutions to 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

$$\boxed{x_{LS} = \left\{ x \in \mathbb{C}^n \colon \ \|\mathbf{A}x - b\|_2^2 \ \text{ is minimized} \right\}.} \tag{1.9}$$

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

$$\boxed{\begin{array}{rcccl} x_{LS} & = & \mathbf{A}^{\dagger}b & + & \left(\mathbf{I}_n - \mathbf{A}^{\dagger}\mathbf{A}\right) y, \qquad y \in \mathbb{C}^n \\ & = & x_{\mathcal{R}} & + & x_{\mathcal{N}}, \\ & = & x_{part} & + & x_{homog}. \end{array}} \tag{1.10}$$

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

Chapter 2

Least Squares Theory

Bolstered from producing concrete results, 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 upon which the stories of existence and uniqueness play out.

An exemplar matrix,

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

demonstrates salient points with an immediate subspace decomposition. 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 row space.

For this case, the domain can be described as

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

and the codomain as

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

The symbol $\oplus$ specifies that the vectors spaces for the range and null space are orthogonal.

There are many statements of the Fundamental Theorem. One way describes the decomposition of the domain and codomain into the four fundamental subspaces: Each domain space, domain and codomain, are expressed as an orthogonal

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

domain:	$\mathbb{C}^n$	$=$	$\mathcal{R}(\mathbf{A}^*)$	$\oplus$	$\mathcal{N}(\mathbf{A})$
codomain:	$\mathbb{C}^m$	$=$	$\mathcal{R}(\mathbf{A})$	$\oplus$	$\mathcal{N}(\mathbf{A}^*)$

decomposition of a range space and a null space.

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

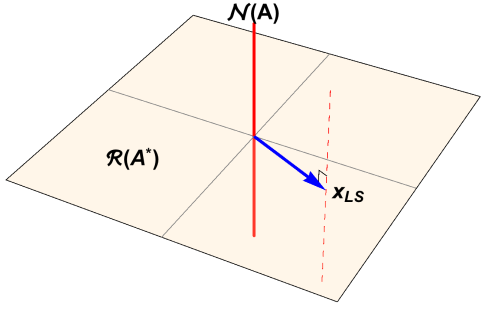
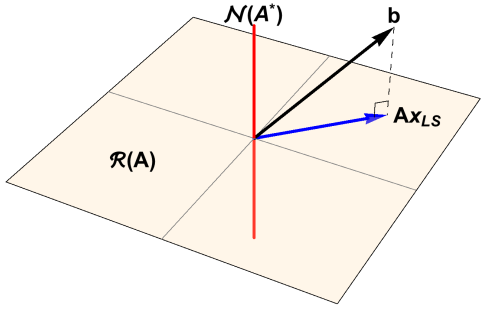
Domain	Mapping	Codomain
	$\mathbf{A}\colon \mathbb{C}^n \mapsto \mathbb{C}^m$ $\mathbb{C}^n \mapsto \mathbb{C}^m\colon \mathbf{A}^*$	
$\mathbb{C}^n = \mathcal{R}(\mathbf{A}^*) \oplus \mathcal{N}(\mathbf{A})$		$\mathbb{C}^m = \mathcal{R}(\mathbf{A}) \oplus \mathcal{N}(\mathbf{A}^*)$

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:

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

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

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

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

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

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

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. As noted by Gene Golub, the singular value decomposition is a singularly valuable decomposition. It not only finds an orthonormal basis for the fundamental subspaces, it also aligns the domain and codomain, and furthermore, resolves the scale factors needed to move between spaces. This first presentation offers a statement of the SVD theorem.

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

$$\mathbf{A} = \mathbf{U}\,\Sigma\,\mathbf{V}^* \tag{2.1}$$

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 (1.9) 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:

$$\|\mathbf{A}x - b\|_2^2 = \|\mathbf{U}^* (\mathbf{A}x - b)\|_2^2 . \tag{2.2}$$

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

$$\mathbf{S}\mathbf{V}_{\mathcal{R}}^{*}x - \mathbf{U}_{\mathcal{R}}^{*}b. \tag{2.3}$$

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

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

This is also the pseudoinverse solution

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

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

We can reimagine the least squares problem as the challenge of resolving the data vector into range and null space components:

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

Such an example is shown in (5.2).

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

Because the vector $b_{\mathcal{R}} \in \mathcal{R}(\mathbf{A})$, there exists a vector x such that $\mathbf{A}x = b_{\mathcal{R}}$. Again, the error that cannot be removed is the residual error

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

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

$$r^2 = \|\mathbf{U}_{\mathcal{N}}^{*}b\|_2^2 = \left(\mathbf{U}_{\mathcal{N}}^{*}b\right)^{*}\left(\mathbf{U}_{\mathcal{N}}^{*}b\right) = b^{*}\left(\mathbf{U}_{\mathcal{N}}\mathbf{U}_{\mathcal{N}}^{*}\right)b, \tag{2.4}$$

a restatement of (2.2.2).

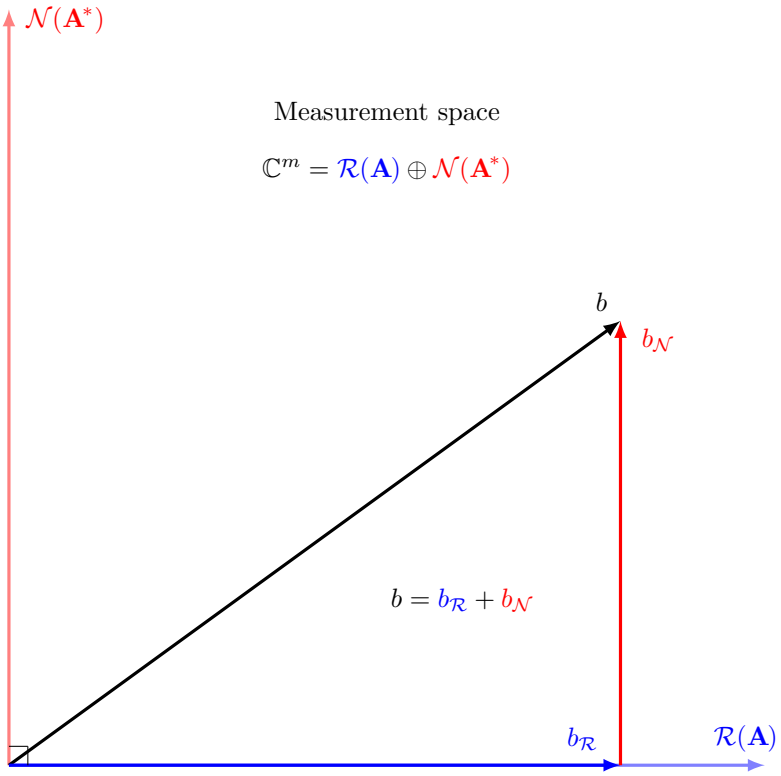


Figure 2.1. *Decomposing the data vector.*

2.3 Singular Value Decomposition - II

We have a statement of the singular value decomposition in §2.2.1. We have an understanding that the SVD revolves the four fundamental subspaces and the scaling factors between range spaces (eq. (2.1)). We saw in §2.2.2 how the SVD naturally leads to the pseudoinverse solution for the least squares problem. Now we turn attention to manipulation of the SVD to further explore the least squares method.

2.3.1 Σ Gymnastics

Success in manipulating the singular value decomposition builds upon success in manipulating the Σ matrices. There are three flavors of interest: the raw Σ matrix, the transpose of this matrix, Σ^{T}, and the pseudoinverse $\Sigma^{\dagger}$.

Think of the Σ matrix as a sabot matrix for the matrix of singular values $\mathbf{S}$. The padding of 0's provides compatibility of shape: the domain matrix $\mathbf{U}$ in $m \times m$; the domain matrix $\mathbf{V}$ is $n \times n$. Conformability insists the Σ matrix in the triple product $\mathbf{U}\,\Sigma\,\mathbf{V}^*$ have dimension $m \times n$.

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.

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

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

There are two basic forms of interest:

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	$\mathbf{P}_{\mathcal{R}(\mathbf{A})}$	$=$	$\mathbf{U}_{\mathcal{R}}\mathbb{I}_{m,\rho}\mathbf{U}_{\mathcal{R}}^*$	$\mathbf{P}_{\mathcal{N}(\mathbf{A}^*)}$	$=$	$\mathbf{I}_m - \mathbf{U}_{\mathcal{R}}\mathbb{I}_{m,\rho}\mathbf{U}_{\mathcal{R}}^*$

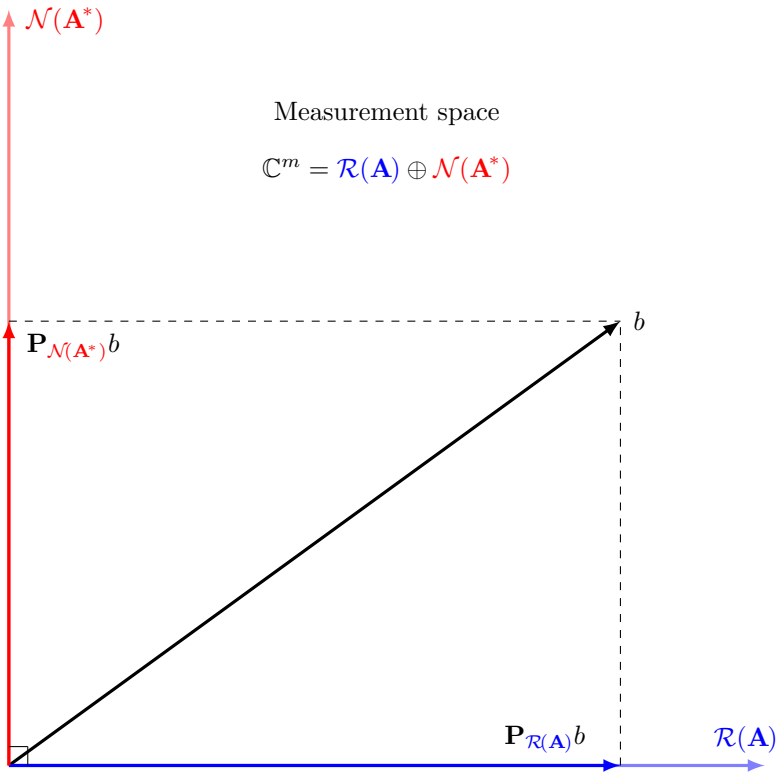


Figure 2.2. *Projections of the data vector.*

2.4 Least Squares Solution - Again

Let's revisit the canonical linear system in (1.1) the general solution in (1.10):

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

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

$$\sum r_k^2 = \sum \left(-r_k\right)^2 .$$

Otherwise,

$$r \cdot r = (-r) \cdot (-r)$$

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

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

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

There are two equivalent ways to define the residual vector. When discussing range and null space decomposition, it is natural to write $b = b_{\mathcal{R}} + b_{\mathcal{N}}$ as in figure 2.3. However, when discussing the projection of the data vector down onto the range space as in figure 2.4, the choice is $b + b_{\mathcal{N}} = b_{\mathcal{R}}$. In the first instance the red arrow points up, in the second it points down.

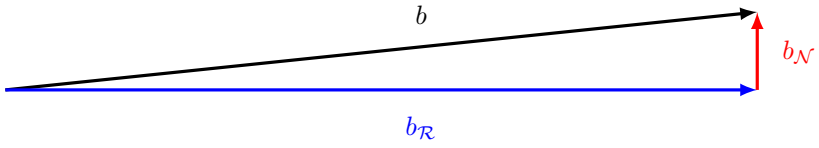


Figure 2.3. *Data vector resolved as* $b = b_{\mathcal{R}} + b_{\mathcal{N}}$.

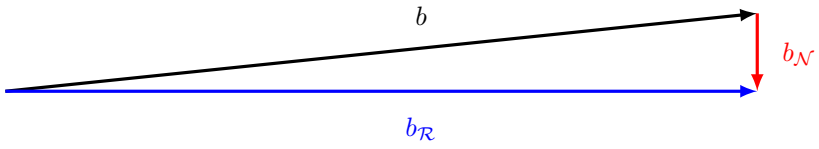


Figure 2.4. *Data vector resolved as* $b + b_{\mathcal{N}} = b_{\mathcal{R}}$.

$$x_{LS} = x_R$$

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

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

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

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

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

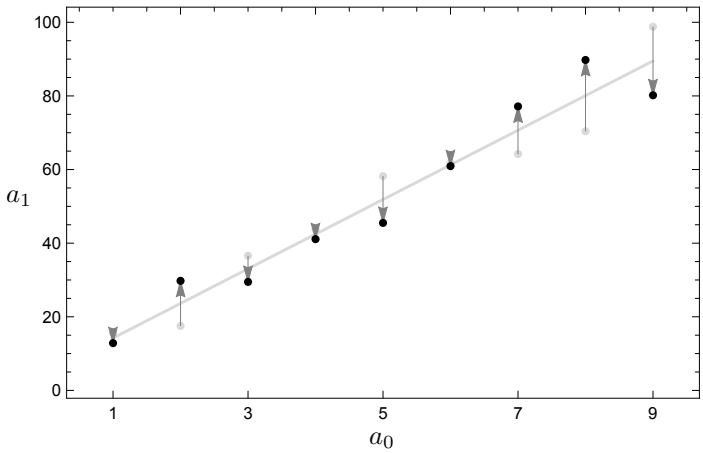


Figure 2.5. *Reflecting the data points through the solution curve.*

This test can be a filter as seen in §14.3.3

Chapter 3

Playing With Lines

The method of least squares is a wonderfully versatile tool, and we will get a sense of this from looking at the canonical equation for a line

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

There are two ways to exploit the equation. One way is to input a set of measurements $\{x_k, y_k\}_{k=1}^{m}$, and find the best parameters (a_0, a_1), in the least squares sense.

The focus on this chapter is to take a set of parameters (a_0, a_1) and to find the bet set (x, y), in the least squares sense.

input	$\rightarrow$	output
parameters a_0, a_1		set (x, y)
intercept, slope		least squares solution

3.1 A Single Line

Given the parameters $a = \begin{bmatrix} a_0 \\ a_1 \end{bmatrix}$, find the set $\begin{bmatrix} x \\ y \end{bmatrix}$ which comprises the least squares solution for

$$y = a_0 + a_1 x, \qquad a_0 \neq 0. \tag{3.1}$$

Rewriting the equation as a linear system reveals

$$-a_1 x + y = a_0,$$

which is the linear system

$$\begin{bmatrix} -a_1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a_0 \end{bmatrix}.$$

Certainly, the heavy artillery of the SVD could be brought to bear. Yet simple geometric reasoning will produce a solution and reinforce basic insights.

Equation (3.1) describes a line, a continuous set of points. Referring back to the boxed least squares solution in (1.10), the least squares solution may be the

Which point on the line is closet to the origin? One way to solve for this is use the fact that the shortest line through the origin which connects with (3.1) is perpendicular to (3.1). This is the orthogonal projection theorem.

Table 3.1. *Problem statement for least squares solution for a single line.*

trial function	$y(x) = a_0 + a_1 x$	
residual error	$r(x) = y - a_0 - a_1 x$	
merit function	$M(a) = \sum_{k=1}^{m} \left(T_k - a_0 - a_1 x\right)^2$	
inputs	a_0	$y-$axis intercept
	a_1	slope
results	$\begin{bmatrix} x \\ y \end{bmatrix}_{LS} = \begin{bmatrix} x \\ y \end{bmatrix}_{part} + \begin{bmatrix} x \\ y \end{bmatrix}_{homog}$	
# of parameter sets	$m = 1$	rows in $\mathbf{A}$
# of parameters	$n = 2$	columns in $\mathbf{A}$
system matrix	$\mathbf{A} \in \mathbb{R}_1^{1\times 2}$	
linear system	$\begin{bmatrix} -a_1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a_0 \end{bmatrix}$	
ideal solution	NA	
input data	–	

$$\begin{bmatrix} -a_1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a_0 \end{bmatrix}$$

$$x_{LS} = \left\{ x \in \mathbb{R}^2 \colon \left\| \mathbf{A}x - a_0 \right\|^2 \text{ is minimized} \right\}$$

$$M\left(x, y\right) = \left(y - a_0 - a_1 x\right)^2$$

Because every point on the line (3.1) is a point of 0 residual, every point on the line is a solution point.

The least squares solution is the solution of minimum length.

$$y = -\frac{1}{a_1} x$$

$$x = \frac{a_0 a_1}{1 + a_1^2}$$

$$p(\tau) = p_0 + \tau \begin{bmatrix} 1 \\ a_1 \end{bmatrix}$$

$$\mathbf{A}^\dagger x = \frac{a_0}{1 + a_1^2} \begin{bmatrix} -a_1 \\ 1 \end{bmatrix}$$

Every point on the line in (3.1)

3.2 Two Lines

Where Do Parallel Lines Cross? The provocative question which opens this section has an obvious answer in Euclidean space. There is no such point; parallel lines never cross. Yet, if we input these lines as a linear system, we compute a least squares solution. What is the significance of the least squares solution?

Consider this tantalizing example. The two lines,

$$y(x) = \tfrac{1}{2} - \tfrac{1}{2}x,$$
$$y(x) = 1 - \tfrac{1}{2}x,$$

are plotted in figure 3.1 along with the least squares solution is

$$x_{LS} = \mathbf{A}^\dagger b = \frac{1}{10} \begin{bmatrix} 3 \\ 6 \end{bmatrix}.$$

What is so special about this point?

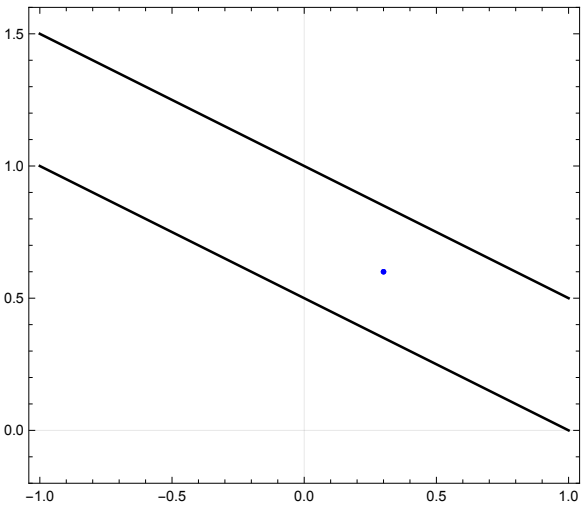


Figure 3.1. *Parallel lines and the least squares solution.*

3.2.1 Intersecting Lines

$$y = 1,$$
$$y = x.$$

$$\begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

3.3 Three Lines

$$y_1(x) = 1,$$
$$y_2(x) = 1 - x,$$
$$y_3(x) = mx.$$

Table 3.2. *Rewriting the equations $y = mx + b$ as a linear system.*

$$\begin{array}{lcllrcrcl} y_1(x) & = & 1 & \Rightarrow & & & y & = & 1 \\ y_2(x) & = & 1 - x & \Rightarrow & x & + & y & = & 1 \\ y_3(x) & = & mx & \Rightarrow & -mx & + & y & = & 0 \end{array}$$

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \\ -m & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$$

Merit function

$$M(x, y) = \left\| \mathbf{A} \begin{bmatrix} x \\ y \end{bmatrix} - b \right\|_2^2$$

$$\mathbf{A}^* \cdot \mathbf{A} = \begin{bmatrix} 1 + m^2 & 1 - m \\ 1 - m & 3 \end{bmatrix}$$

$$\operatorname{tr}(\mathbf{A}) = 4 + m^2$$
$$\det(\mathbf{A}) = 2 + 2m + 2m^2$$

$$p(\lambda) = \lambda^2 - \lambda \operatorname{tr}(\mathbf{W}) + \det(\mathbf{W})$$
$$p(\lambda) = 0$$

$$\lambda_{\pm} = \frac{\operatorname{tr}(\mathbf{W}) \pm \sqrt{\operatorname{tr}(\mathbf{W})^2 - 4 \det(\mathbf{W})}}{2}$$

$$\sigma = 2^{-1/2}\sqrt{m^2 + 4 \pm \sqrt{m^4 - 8m + 8}}$$

$$\xi = \sqrt{m^4 - 8m + 8}$$

$$p(m) = \frac{1}{2\left(m^2+m+1\right)} \begin{bmatrix} b_1(m-1) + b_2(m+2) + b_3(-2m-1) \\ b_1\left(m^2+1\right) + b_2 m(m+1) + b_3(m+1) \end{bmatrix} \tag{3.2}$$

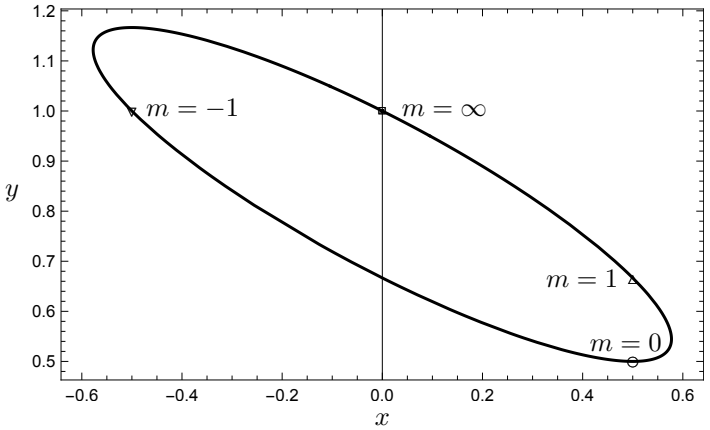


Figure 3.2. *Trajectory of the solution point* $p(m)$ *in* (3.2) *for* $-\infty < m < \infty$.

Table 3.3. *Least squares solution for three distinct lines as the parameter m varies from 0 to ∞.*

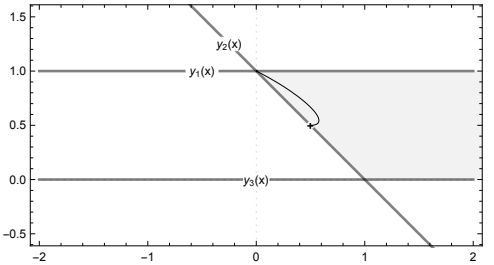
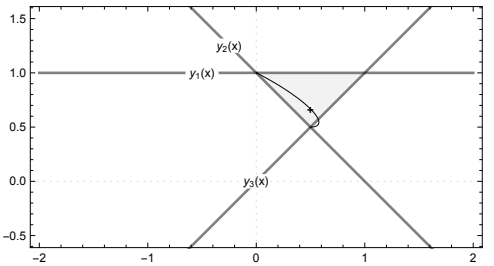
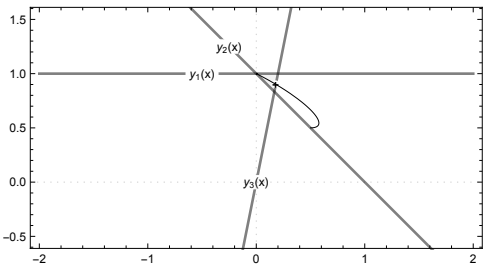
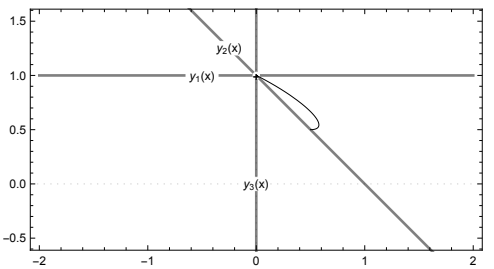
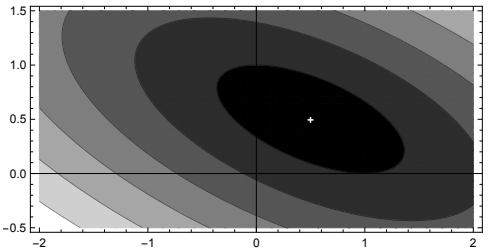
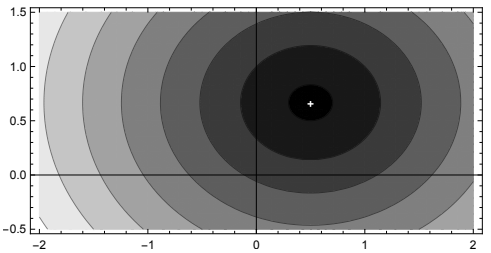
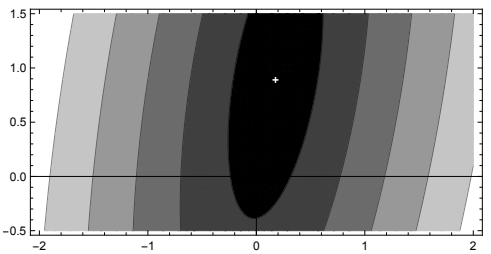
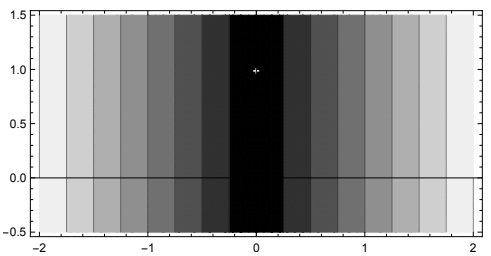
m	Domain: Graphs
$m = 0$	
$m = 1$	
$m = 5$	
$m = \infty$	

Table 3.4. *Least squares solution for three distinct lines as the parameter m varies from 0 to ∞.*

m	Domain: Merit function
$m = 0$	
$m = 1$	
$m = 5$	
$m = \infty$	

Part II

Modal Example

A sample problem examined from a few different perspectives.

Chapter 4

Solution via Calculus

4.1 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 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,

$$\begin{aligned} M(a) &= \sum_{k=1}^{m} r_k^2 \\ &= \sum_{k=1}^{m} \left(\text{measurement}_k - \text{prediction}_k\right)^2 \\ &= \sum_{k=1}^{m} \left(T_k - T\left(x_k\right)\right)^2 \\ &= \sum_{k=1}^{m} \left(T_k - a_0 - a_1 x_k\right)^2 \end{aligned} \tag{4.1}$$

The least squares solution a_{LS} is defined as

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

Colloquially, the least squares solution is the set a of complex pairs such that the square of the norm $\|T(x_k) - a_0 - a_1 x_k\|_2^2$ is minimized.

The solution set satisfies

$$\nabla M(a)|_{a_{LS}} = 0. \tag{4.2}$$

4.2 Bevington Example

To provide a common reference, see the example in Bevington [2, ch 6]. The data is summarized below in table 4.2. 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 4.1. 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".

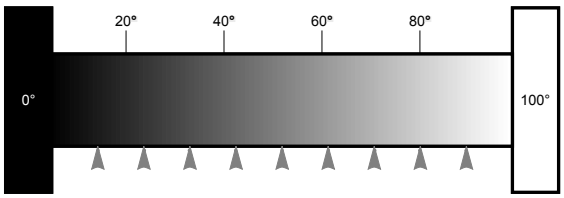


Figure 4.1. *Measuring the temperature of a bar held between two constant-temperature heat baths.*

4.2.1 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 14.4.

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 (4.1). 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 4.2. Discussion of significant digits in the input data is deferred.

4.2.2 Normal Equations via Calculus

In §6.4, Bevington solves the problem by applying calculus to the final form in (4.1), effectively solving (4.2). 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

$$\begin{aligned} \partial_0 M(a_0, a_1) &= 0, \\ \partial_1 M(a_0, a_1) &= 0. \end{aligned} \tag{4.3}$$

These two differential equations spawn the linear system

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

Table 4.1. *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 **A**
# of parameters	$n = 2$	columns in **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 4.2	

Distributing the summation operators creates a more revealing form

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

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

$$\begin{bmatrix} \sum 1 & \sum x_k \\ \sum x_k & \sum x_k^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum T_k \\ \sum T_k x_k \end{bmatrix}. \tag{4.4}$$

The solution can be written immediately. Defining the determinant

$$\Delta = m \sum x_k^2 - \left(\sum x_k \right)^2, \tag{4.5}$$

the matrix inverse is

$$\begin{bmatrix} m & \sum x_k \\ \sum x_k & \sum x_k^2 \end{bmatrix}^{-1} = \Delta^{-1} \begin{bmatrix} \sum x_k^2 & -\sum x_k \\ -\sum x_k & m \end{bmatrix}. \tag{4.6}$$

Table 4.2. *Raw data and results.*

	Input		Output	
	(cm)	(°C)	(°C)	(°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 (4.4) is the matrix product

$$\begin{aligned}
\begin{bmatrix} a_0 \\ a_1 \end{bmatrix} &= \Delta^{-1} \begin{bmatrix} \sum x_k^2 & -\sum x_k \\ -\sum x_k & \sum 1 \end{bmatrix} \begin{bmatrix} \sum T_k \\ \sum T_k x_k \end{bmatrix}, \\
a_0 &= \Delta^{-1} \left(\sum x_k^2 \sum T_k - \sum x_k \sum T_k x_k \right), \\
a_1 &= \Delta^{-1} \left(m \sum T_k x_k - \sum x_k \sum T_k \right).
\end{aligned} \tag{4.7}$$

The final results agree with Bevington's equations 6–17.

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 (4.6):

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

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

Table 4.3. *Results for linear regression.*

fit parameters	$a_0 = \frac{\sum x_k^2 \sum x_k y_k - \sum x_k \sum x_k y_k}{m \sum x_k^2 - (\sum x_k)^2}$	intercept, °C
	$a_1 = \frac{m \sum x_k y_k - \sum x_k \sum y_k}{m \sum x_k^2 - (\sum x_k)^2}$	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 14.4	
input data	table 4.2	
plots	figure 4.2	1. data & solution
	figure 4.3	2. residual errors
	figure **??**	3. merit function

4.3 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.

4.3.1 Exact Form

Exact results for the fit parameters are error follow. The product matrix in (4.4) 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 (9.3)

$$\Delta = 540.$$

The inverse of this matrix, (4.6), 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, (4.7), 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

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

4.3.2 Computed Form

The results in the previous section are shown in integer form as debugging tool: readers can check answer to arbitrary precision. 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:

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

Quote the total error as $r^*r \approx 320$.

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

$$\begin{aligned} \epsilon_0 &= 4.886206312183354 && \text{rounded to } 4.9, \\ a_0 &= 4.813888888888889 && \text{rounded to } 4.8; \\ \epsilon_1 &= 0.8683016476563611 && \text{rounded to } 0.87, \\ a_1 &= 9.408333333333333 && \text{rounded to } 9.41. \end{aligned}$$

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.

4.4 Visualization

4.4.1 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.

Data v. Fit

The first plot is the solution curve plotted against the measurements as shown in figure 4.3. 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 §5.2.

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.

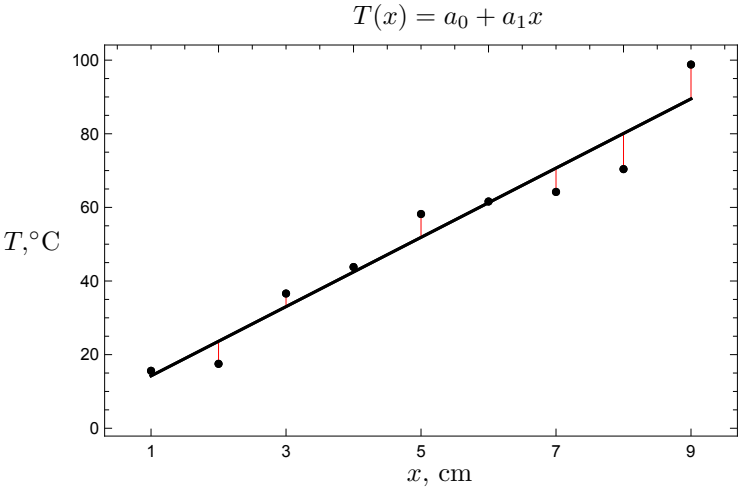


Figure 4.2. *Solution plotted against data with residual errors shown in red.*

Residual Errors

Next, examine the residual errors as in figure 4.3. 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. A lone point five standard deviations from the mean stands out here. Yes, it will also be apparent on the solution plot, but that does not show rejection criteria in terms of the standard deviation.

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 4.4. The result should be a jittery line with an oscillating appearance, not a trend line. As in this case.

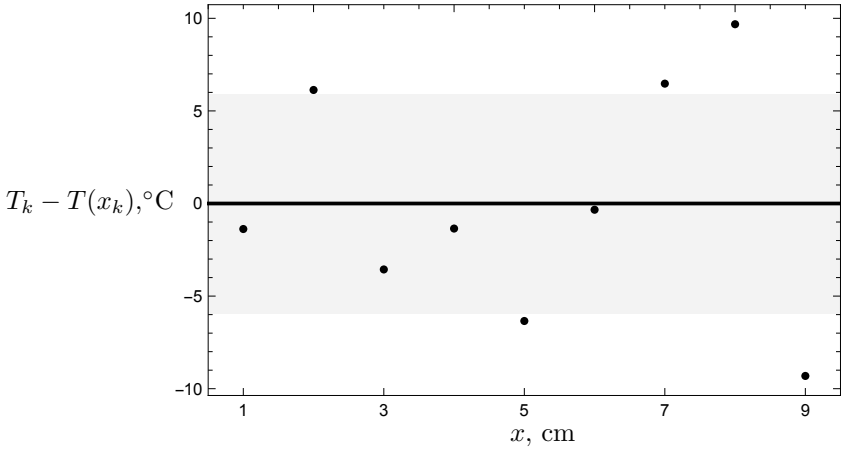


Figure 4.3. *Scatter plot of residual errors showing mean (thick line) and one standard deviation (gray box).*

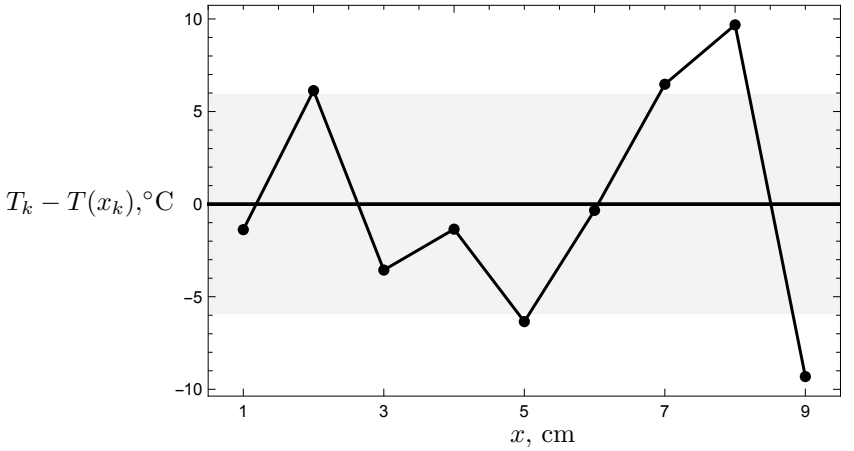


Figure 4.4. *Scatter plot of residual errors with data points connected.*

Solution v. Minimum

The third and final diagnostic plot, the bullseye plot, figure 4.4.1, 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 4.4.1.

The modified version explicitly labels the radii of the error ellipse, ϵ_0, and

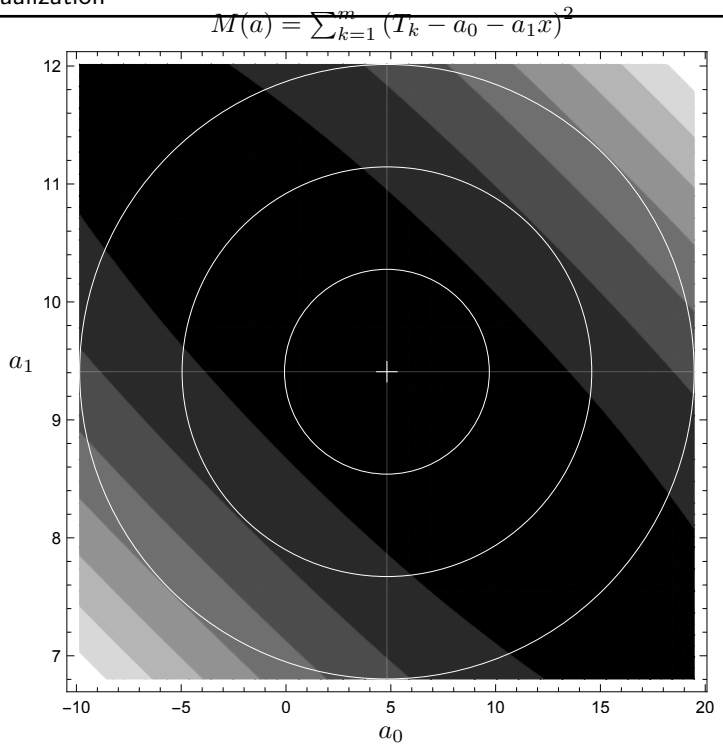


Figure 4.5. *Contour plot of merit function showing the solution (white cross) and three concentric error bands.*

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.

4.4.2 Seeing the Uncertainty

The interpretation of the variation of the parameters a is straightforward. A nice visualization is the "whisker plot" in figure 4.7. 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.

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

Each solution represents a graph. All of the graphs provide a qualitative feel for the simultaneous variation of slope and intercept prescribed by the solution. The white dots representing the ideal solution.

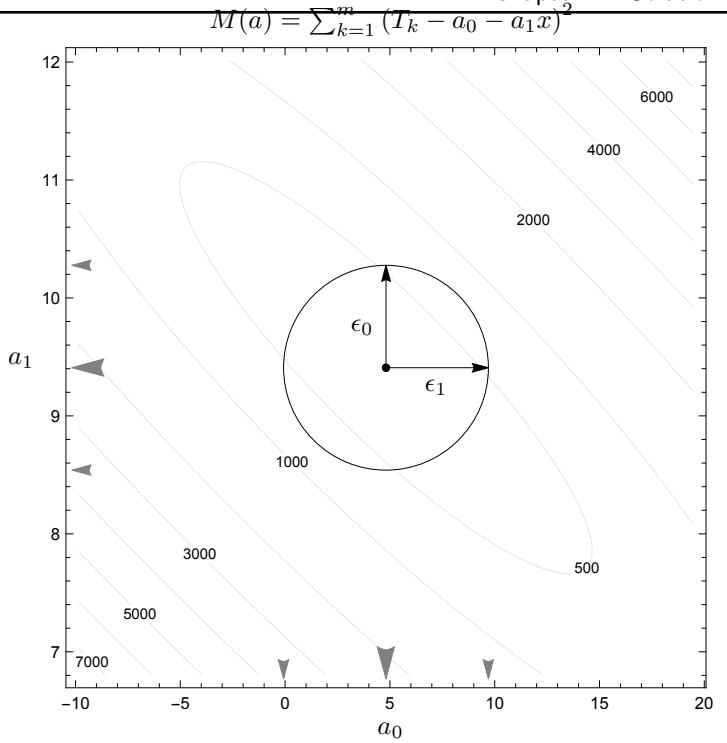


Figure 4.6. *Another look at the merit function showing the primary error ellipse (black) and contour levels (gray).*

4.4.3 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 foundation is the assumed distribution of errors, here the normal distribution:

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

The curve is normalized to unity, that is,

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

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 4.4 shows these connections explicitly and provides a qualitative representation for the quality of the parameters.

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.

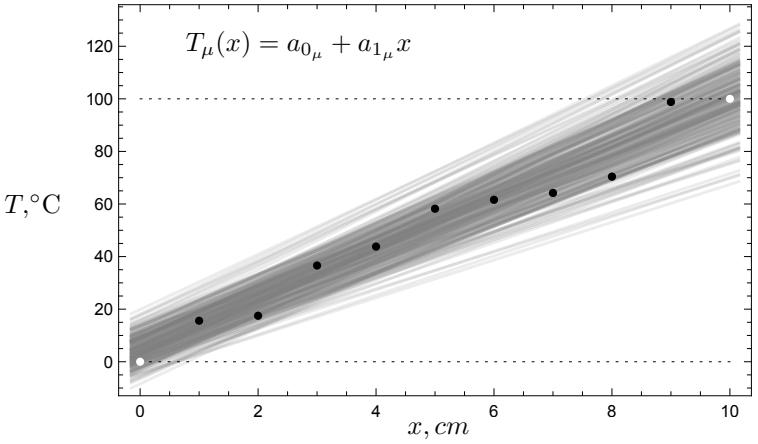
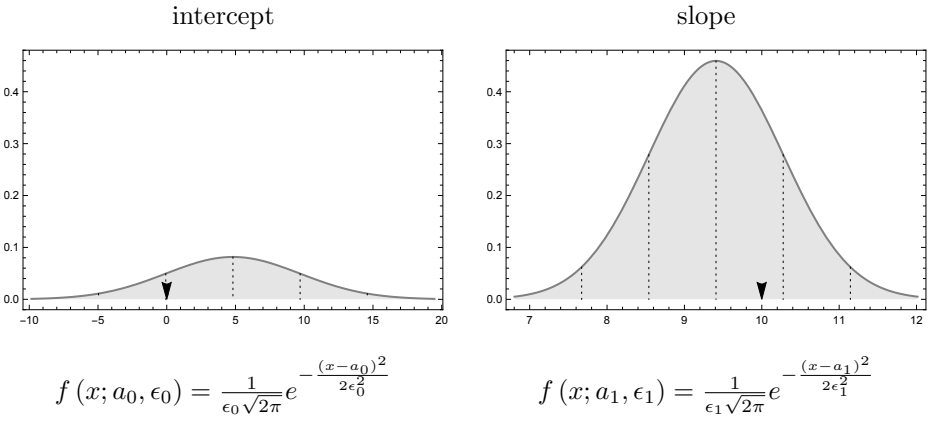


Figure 4.7. *Whisker plot showing 250 randomly sampled solutions.*

Table 4.4. *The solution parameters expressed as normal distributions.*



Arrowheads locate the ideal solution and dotted lines demarcate the first three standard deviations. This provides a visual gauge of the accuracy and precision of the calculation.

Figure 4.8 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 4.5 displays the actual census and compares to the continuum results.

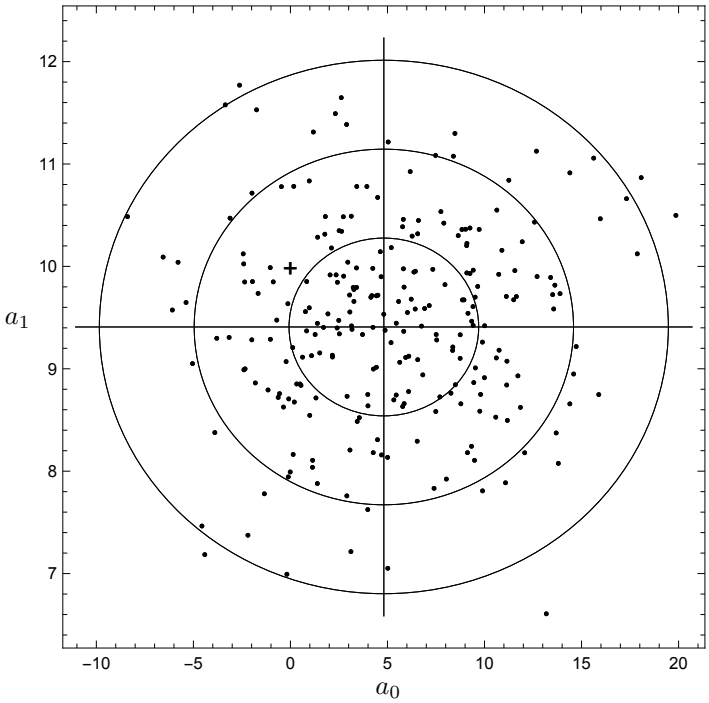


Figure 4.8. *Scatter plot showing sampling of 250 solution points in* $\mathcal{R}(\boldsymbol{A}^*)$.

Table 4.5. *Comparing samples to ideal normal distribution.*

annulus	count	rel. area	count density	normalized density	cumulative density	continuum 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 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. These normalized densities approximate the final column, the continuum limit given by

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

Chapter 5

Solution via SVD

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

5.1 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

$$\mathbf{S} = \left[\begin{array}{cc} \sigma_1 & 0 \\ 0 & \sigma_2 \end{array}\right].$$

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).

5.1.1 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 (9.3) and (6.2); 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

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

(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 $\mathbf{S}$:

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

5.1.2 Matrix for the domain

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

$$\Sigma^{\mathrm{T}}\Sigma = \left[\begin{array}{cc} \mathbf{S} & \mathbf{0} \end{array}\right] \left[\begin{array}{c} \mathbf{S} \\ \mathbf{0} \end{array}\right] = \mathbf{S}^2$$

$$\mathbf{V}_{\mathcal{R}}\mathbf{S}^2\mathbf{V}_{\mathcal{R}}^* = \mathbf{A}^*\mathbf{A}. \tag{5.1}$$

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

$$\mathbf{V}_{\mathcal{R}} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{5.2}$$

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 (**??**), (5.1), and (5.2) 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

$$\cos\theta = \sqrt{\frac{\sigma_2^2 - \mathbf{1}^{\mathrm{T}}\mathbf{1}}{\sigma_2^2-\sigma_1^2}} = \sqrt{\frac{x^{\mathrm{T}}x - \sigma_1^2}{\sigma_2^2-\sigma_1^2}}. \tag{5.3}$$

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

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

5.1.3 Matrix for the codomain

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.

5.1.4 Error terms

The error terms can be computed after this step. Given that the product matrix decomposition in (5.1) is a singular value decomposition we can trivially write the inverse matrix as

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

The diagonal elements describe the connection between error in the data and error in the solution parameters

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

Using the columns vectors and the singular values, the error terms in (5.1.4) become

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

$$= \sqrt{\frac{r^* r}{m - n}} \sqrt{\begin{bmatrix} \sigma_1^2 - \sqrt{(\sigma_2^2 - \mathbf{1}^{\mathrm{T}}\mathbf{1})(\sigma_2^2 - \sigma_1^2)} \\ \sigma_2^2 + \sqrt{(\sigma_2^2 - \mathbf{1}^{\mathrm{T}}\mathbf{1})(\sigma_2^2 - \sigma_1^2)} \end{bmatrix}}.$$

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

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

5.2 Visualization

With the singular value decomposition in hand, the domain space plots become more concrete and we do so below beginning in figure (5.2). 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

$$\begin{aligned}\begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} &= \left(30\left(\sqrt{2341}-31\right)\sqrt{4682}+58\sqrt{2341}\right)^{-1} \times \\ &\qquad \begin{bmatrix} 4\,104\,889+75\,341\sqrt{2341} \\ 3\sqrt{15}\left(753\,593-15\,933\sqrt{2341}\right) \end{bmatrix} \\ &\approx \begin{bmatrix} 171.733 \\ -4.45594 \end{bmatrix}\end{aligned}$$

$$T_{\mathcal{R}} \approx 171.733u_1 - 4.45594u_2 \tag{5.4}$$

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 (5.2) and displayed in figure 5.1.

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

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

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

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

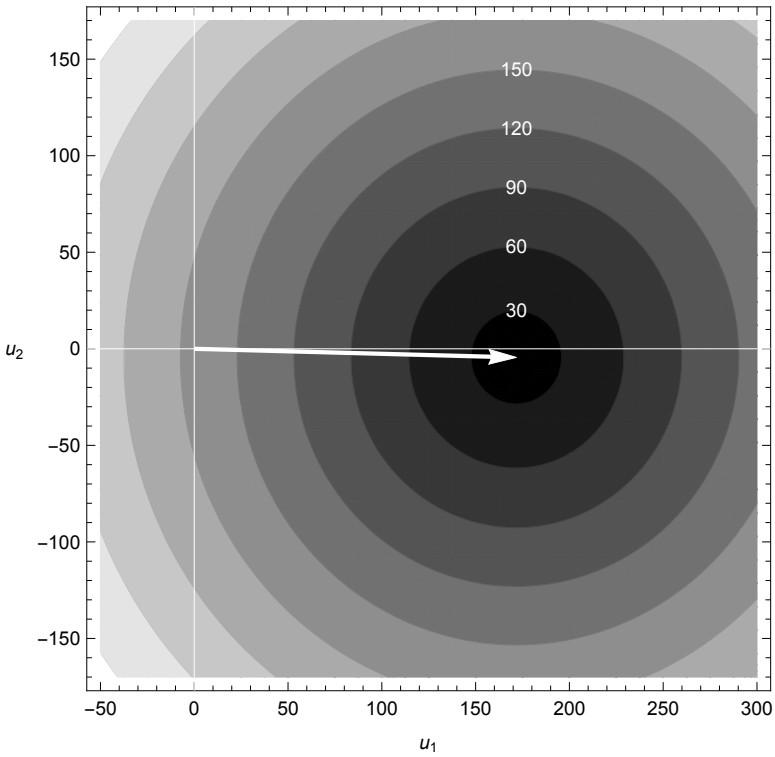


Figure 5.1. *The solution vector in* (5.4) *(white arrow) is the mixture of* u_1 *and* u_2 *which eliminates the error.*

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 5.2 shows an singular value decomposition in block form. The gray values in the system matrix $\mathbf{A}$ shows the values of the matrix entries bound between a low of 1 and a high of 9. The gray shading demonsrates the fact that the column

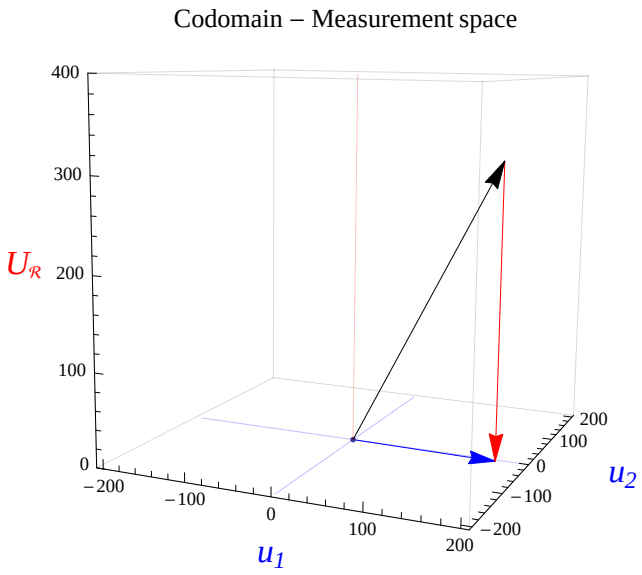


Figure 5.2. *Measurement space $\mathcal{R}(\boldsymbol{A})$ for the Bevington example showing resolution of the data vector (black) into range (blue) and null space (red) components.*

vectors are not in $\mathcal{R}(\mathbf{A})$ and the row vectors are not in $\mathcal{R}(\mathbf{A}^*)$. Because the domain matrices $\mathbf{U}$ and $\mathbf{V}$ are unitary, the magnitude of each entry is less than unity. The matrix for the domain, $\mathbf{U}$, has column vectors in $\mathcal{R}(\mathbf{A})$ shaded in blue, and column vectors in $\mathcal{N}(\mathbf{A})$ shaded in red.

The decomposition shown in figure 2.1.

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

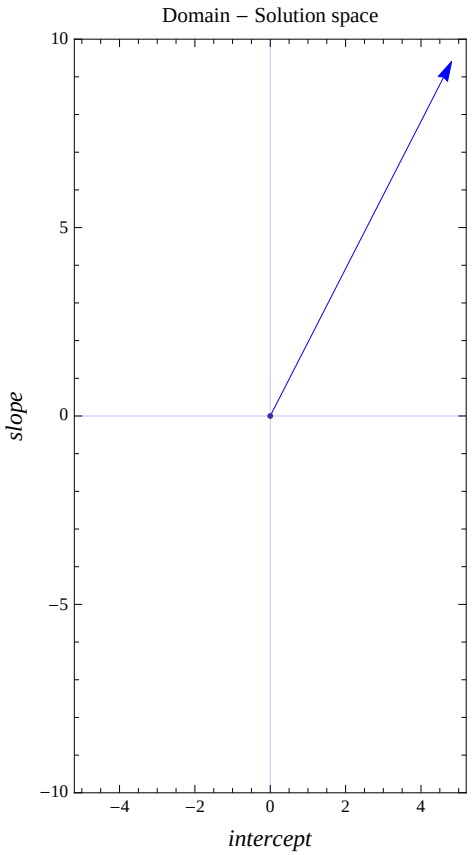


Figure 5.3. *Minimization occurs in the codomain.*

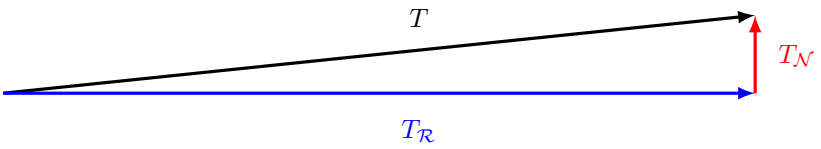


Figure 5.4. *Data vector $T = T_{\mathcal{R}} + T_{\mathcal{N}}$ resolved into range and null space components as in figure 2.3.*

Figure 5.5. *Decomposing $\|r = T_{\mathcal{N}}\|_2^2$ into residual error terms r_k^2 of table 5.3.*

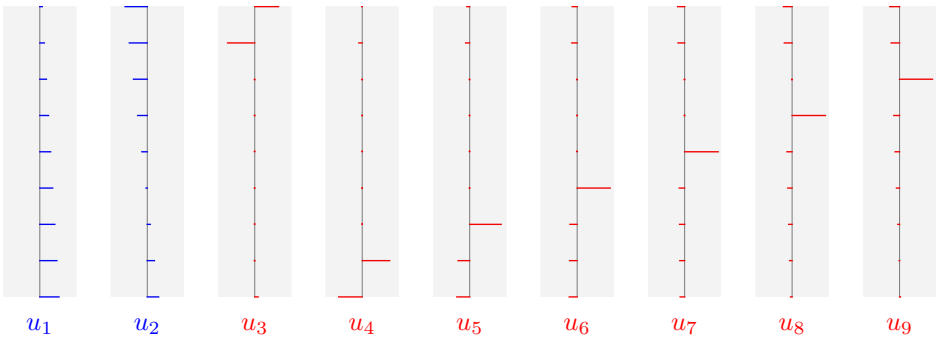


Table 5.1. *The column vectors of* $\mathbf{U}$*. The gray box represents the maximum length an element may have:* $[-1, 1]$*.*

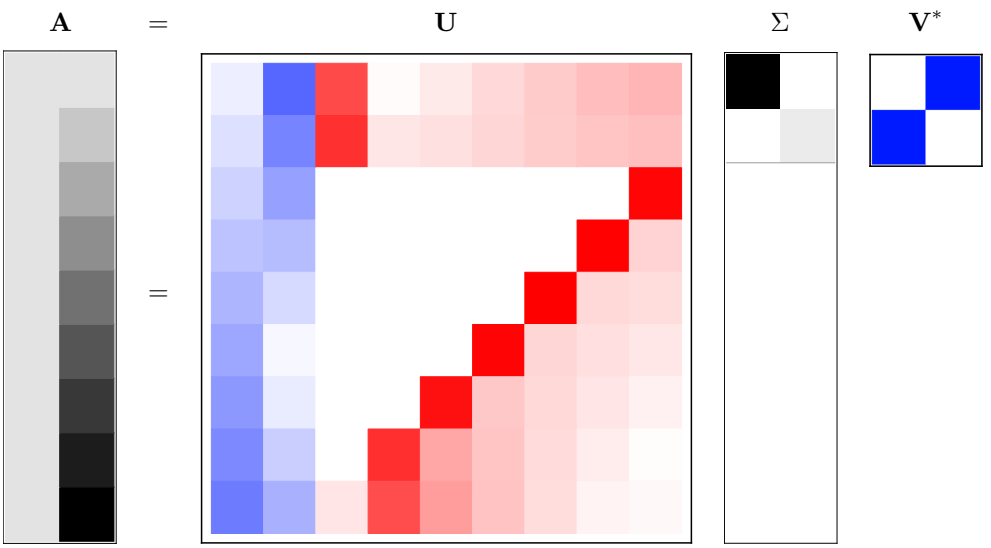


Table 5.2. *Singular value decomposition for the system matrix* $\boldsymbol{A}$ *in* (6.1) *showing range and null space components.*

Table 5.3. *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

Chapter 6

Solution Via Other Methods

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

6.1 Normal Equations from Vectors

In §4.2.2, the normal equations appeared from the linear system resulting from the two partial differential equations in (4.3). 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 4.2. 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:

$$\begin{array}{ccccc} \mathbf{A} & & a & = & T \\ \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \\ 1 & 7 \\ 1 & 8 \\ 1 & 9 \end{bmatrix} & & \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} & = & \frac{1}{10} \begin{bmatrix} 156 \\ 175 \\ 366 \\ 438 \\ 582 \\ 616 \\ 642 \\ 704 \\ 988 \end{bmatrix} . \end{array} \tag{6.1}$$

6.1.1 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

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

Therefore there is no exact solution a such that

$$\|\mathbf{A}a - T\|_2^2 = 0.$$

The root cause is that the data vector has components in the range space and in the null space:

$$T = T_{\mathcal{R}} + T_{\mathcal{N}}.$$

One strategy is to compose a problem 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

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

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,

$$\Delta = \left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)\left(x^{\mathrm{T}}x\right) - \left(\mathbf{1}^{\mathrm{T}}x\right)^{2}, \tag{6.2}$$

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

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

$$\begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \Delta^{-1}\begin{bmatrix} \left(x^{\mathrm{T}}x\right)\left(\mathbf{1}^{\mathrm{T}}T\right) - \left(\mathbf{1}^{\mathrm{T}}x\right)\left(x^{\mathrm{T}}T\right) \\ \left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)\left(x^{\mathrm{T}}T\right) - \left(\mathbf{1}^{\mathrm{T}}x\right)\left(\mathbf{1}^{\mathrm{T}}T\right) \end{bmatrix} \tag{6.3}$$

leading to the same solution presented in §4.3.1.

6.2 QR Decomposition

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

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

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

6.2.1 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:

$$r_{11} = \|a_1\|_2 = 3 \tag{6.4}$$

Normalized column vector

$$q_1 = \frac{a_1}{r_{11}} = \frac{1}{3} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \tag{6.5}$$

Next scale factor in row 1:

$$r_{12} = q_1^* a_2 = 15 \tag{6.6}$$

$$q_2 = a_2 - r_{12} q_1 \tag{6.7}$$

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

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

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

Chapter 7

Finer Points

A deeper look at the example.

7.1 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.

7.1.1 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 4.1, 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 shifts 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

$$x \to x + \mathbf{1}x_*. \tag{7.1}$$

Demonstrations

First compute the transformation rules on the inner products using (7.1):

$$\begin{aligned} &\mathbf{1}^{\mathrm{T}}\mathbf{1} \to \mathbf{1}^{\mathrm{T}}\mathbf{1}, \\ &\mathbf{1}^{\mathrm{T}}x \to \mathbf{1}^{\mathrm{T}}x + x_* \left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right), \\ &x^{\mathrm{T}}x \to x^{\mathrm{T}}x + 2x_* \left(\mathbf{1}^{\mathrm{T}}x\right) + x_*^2 \left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right), \\ &x^{\mathrm{T}}T \to x^{\mathrm{T}}T + x_* \left(\mathbf{1}^{\mathrm{T}}T\right). \end{aligned} \tag{7.2}$$

Computation of the transformation of the intercept a_0 and slope a_1 requires computation of the transformation of the determinant in (**??**): (6.2):

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

Thus the determinant is invariant under translation. Using the substitutions in (7.2) one can show (exercise ???) the solution parameters transform as expected:

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

Homework: C to F, affine transform $T \to \alpha T + \beta$

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 $x_* = 1$.

$$\begin{aligned} a_0 &\to -\frac{827}{180}, \\ a_1 &\to \frac{1129}{120}. \end{aligned} \tag{7.3}$$

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

In agreement with (7.3).

Visuals

Compare to figure 4.2

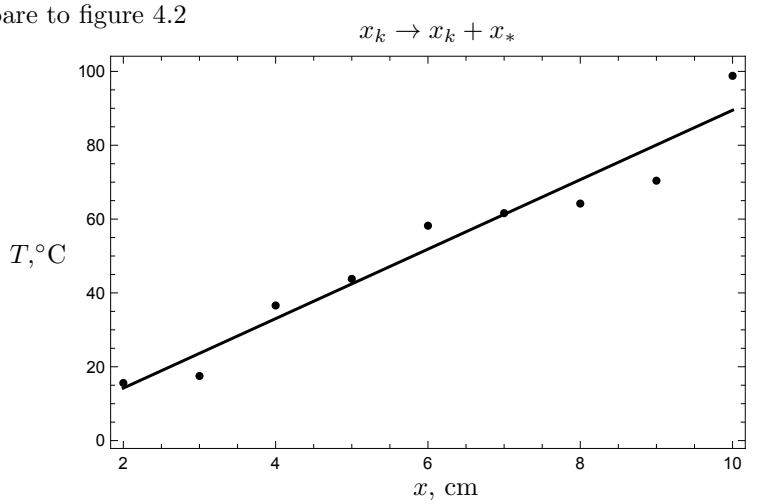


Figure 7.1. *Solution after translation along x axis: the slope is invariant.*

Compare to 4.3
Compare to figure 4.4.1

7.1.2 Reflection Invariance

A demonstration of the reflection method of §2.5.

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

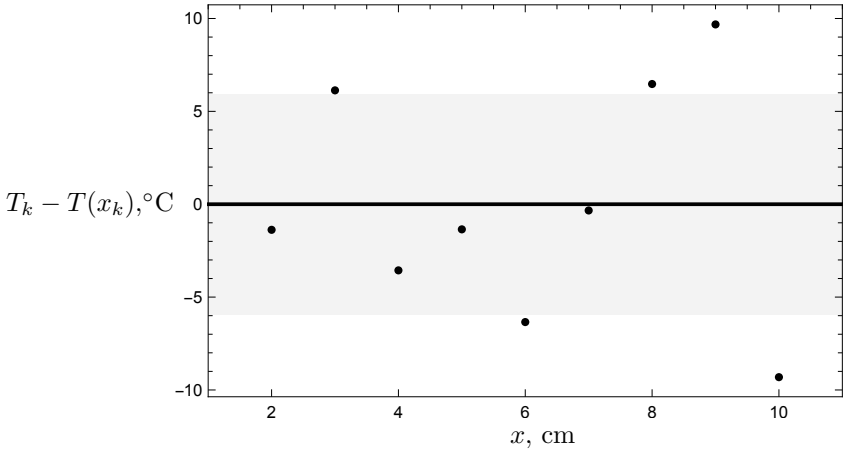


Figure 7.2. *Scatter plot of residual errors after translation along x axis: the residuals are invariant.*

table 14.6

7.2 Fitting To Higher Orders

7.3 Removing Terms

Here we develop a useful tool for removing isolated terms in the merit function.

Trial function for a line (polynomial expansion)

$$y_k = a_0 + a_1 x_k, \qquad k = 1, m$$

In this instance, we can eliminate the intercept by looking at the differences between measurements. To wit,

$$y_j - y_k = a_1 \left(x_j - x_k\right), \qquad \begin{cases} j = 1, m-1 \\ k = j+1, m \end{cases}$$

The common sense interpretation of this equation is that the difference in measurements tells us about the slope Use the integer p_ν, $\nu = \frac{1}{2} n\,(n-1)$ to keep track of the pairs of indices (j, k). The variable $M = \frac{1}{2} n\,(n-1)$ counts the number of distinct pairs.

$$Y_{p_\mu} = a_1 Y_{p_\mu}, \qquad \mu = 1, M$$

The merit function is

$$M(a_1) = \sum_{\mu=1}^{M} \left(Y_{p_\mu} - a_1 X_{p_\mu}\right)^2$$

Setting the derivative with respect to a_1 equal to zero generates the solution

$$a_1 = \frac{\sum_{\mu=1}^{M} X_{p_\mu} Y_{p_\mu}}{\sum_{\mu=1}^{M} X_{p_\mu}^2}$$

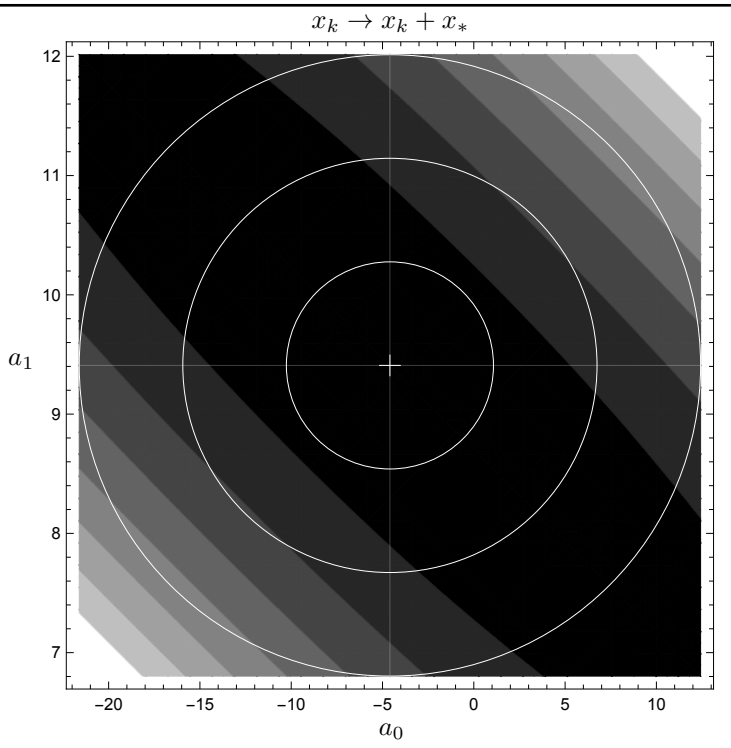


Figure 7.3. *Contour plot of merit function after translation along x axis: a_1 is invariant.*

This is the same answer as (**??**),

$$a_1 = \frac{m\sum_{k=1}^{m} x_k y_k - \sum_{k=1}^{m} x_k \sum_{k=1}^{m} y_k}{m\sum_{k=1}^{m} x_k^2 - \left(\sum_{k=1}^{m} x_k\right)^2}.$$

Numerator:

$$\sum_{j=1}^{m-1}\sum_{k=j+1}^{m}(x_j y_j + x_k y_k) = (m-1)\sum_{i=1}^{m} x_i y_i$$

$$\sum_{i=1}^{m} x_i \sum_{i=1}^{m} y_i = \sum_{i=1}^{m} x_i y_i - \sum_{j=1}^{m-1}\sum_{k=j+1}^{m}(x_j y_k + x_k y_j)$$

Denominator:

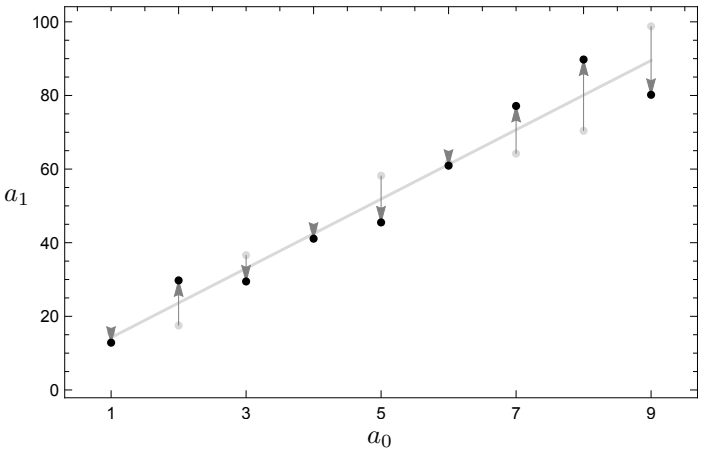


Figure 7.4. *The merit function for the learning curve showing the minimum and the value.*

Table 7.1. *Relating the pair index μ to the indices j and k.*

p_μ	(j,k)
p_1	(1,2)
p_2	(1,3)
$\vdots$	$\vdots$
p_{m-1}	$(1,m-1)$
p_m	$(2,1)$
$\vdots$	$\vdots$
$p_{\frac{1}{2}n(n-1)}$	$(m-1,m)$

Part III

Zonal Example

Chapter 8

Zonal Example

8.1 Problem

8.1.1 Zonal Subsection

Chapter 9

Lines

9.1 Face-centered cubic lattice

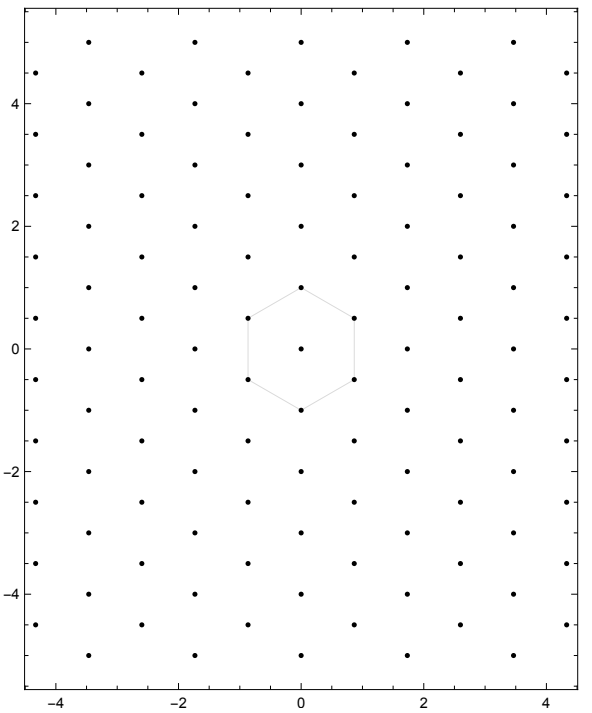


Figure 9.1. *A slice of a face-centered cubic lattice showing a single crystal.*

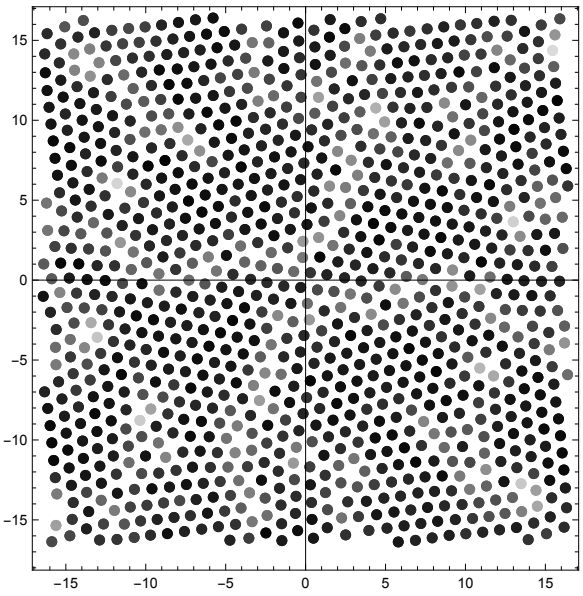


Figure 9.2. *Simulation output showing atomic shades shaded by potential energy.*

9.2 Model

$$y_{(\mu)}(x) = \mu\alpha_* + \alpha_0 + \alpha_1 x, \qquad \mu = 0, 1, 2, \ldots, M-1. \tag{9.1}$$

$$\begin{cases} 0 \cdot \alpha_* + \alpha_0 + \alpha_1 x_{1_1} = y_{1_1} \\ \qquad\qquad\qquad\qquad \vdots \qquad\qquad row\ 1 \\ 0 \cdot \alpha_* + \alpha_0 + \alpha_1 x_{m_1} = y_{m_1} \end{cases}$$
$$\begin{cases} 1 \cdot \alpha_* + \alpha_0 + \alpha_1 x_{1_2} = y_{1_2} \\ \qquad\qquad\qquad\qquad \vdots \qquad\qquad row\ 2 \\ 1 \cdot \alpha_* + \alpha_0 + \alpha_1 x_{m_2} = y_{m_2} \end{cases} \tag{9.2}$$
$$\begin{cases} 2 \cdot \alpha_* + \alpha_0 + \alpha_1 x_{1_3} = y_{1_3} \\ \qquad\qquad\qquad\qquad \vdots \qquad\qquad row\ 3 \\ 2 \cdot \alpha_* + \alpha_0 + \alpha_1 x_{m_3} = y_{m_3} \end{cases}$$

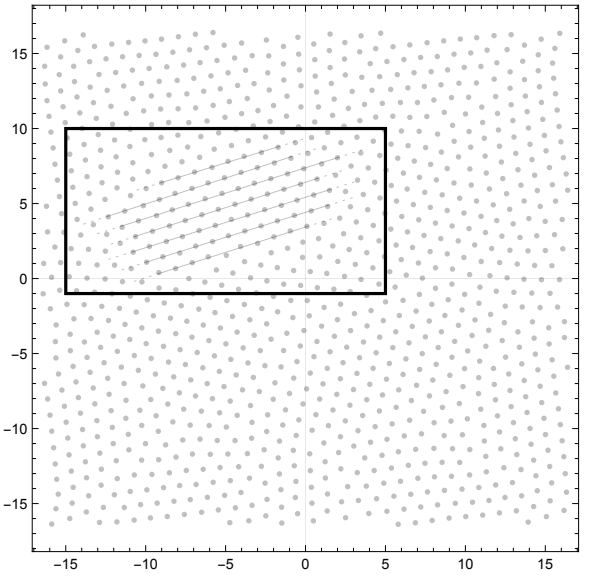


Figure 9.3. *Full data set showing inset.*

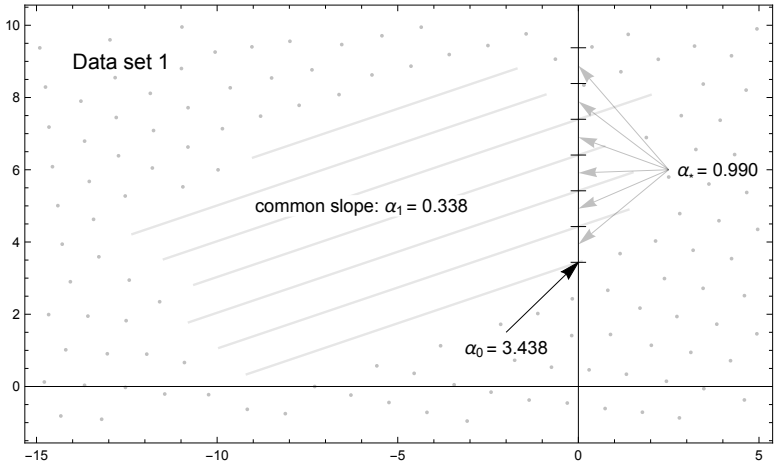
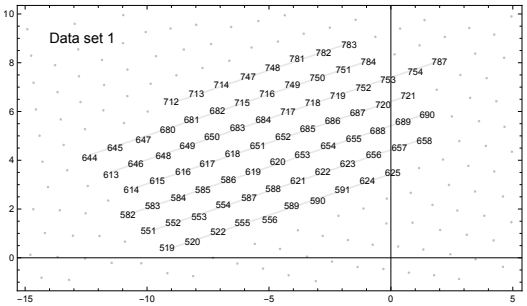
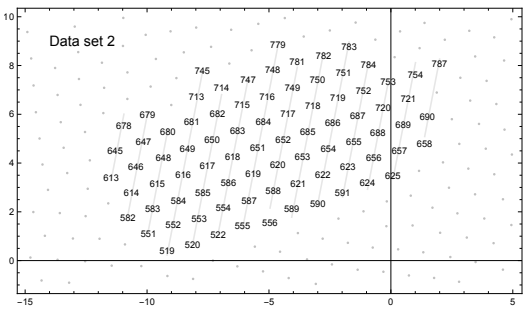


Figure 9.4. *Sample data set showing fit parameters.*

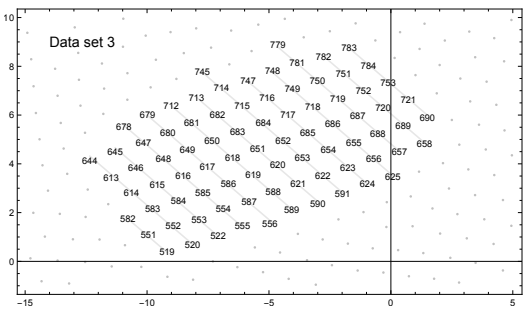
Table 9.1. *Data sets and basic results*



$$\begin{aligned} \alpha_* &= 0.9899 \pm 0.0032 \\ \alpha_0 &= 3.438 \pm 0.013 \\ \alpha_1 &= 0.3376 \pm 0.0017 \\ \sqrt{\langle r^2 \rangle} &= 0.052 \end{aligned}$$



$$\begin{aligned} \alpha_* &= 4.974 \pm 0.052 \\ \alpha_0 &= -2.075 \pm 0.093 \\ \alpha_1 &= 5.168 \pm 0.052 \\ \sqrt{\langle r^2 \rangle} &= 0.18 \end{aligned}$$



$$\begin{aligned} \alpha_* &= 1.2322 \pm 0.0039 \\ \alpha_0 &= -7.505 \pm 0.043 \\ \alpha_1 &= -0.8576 \pm 0.0038 \\ \sqrt{\langle r^2 \rangle} &= 0.054 \end{aligned}$$

$$\begin{array}{ccc} \mathbf{A} & \alpha & y \end{array}$$

$$\left[\begin{array}{ccc} 0 & 1 & x_{1_1} \\ \vdots & \vdots & \vdots \\ 0 & 1 & x_{\mu_1} \\ \hline 1 & 1 & x_{1_2} \\ \vdots & \vdots & \vdots \\ 1 & 1 & x_{\mu_2} \\ \hline 2 & 1 & x_{1_3} \\ \vdots & \vdots & \vdots \\ 2 & 1 & x_{\mu_3} \\ \hline \vdots & \vdots & \vdots \\ \hline M-1 & 1 & x_{1_M} \\ \vdots & \vdots & \vdots \\ M-1 & 1 & x_{m_M} \end{array}\right] \left[\begin{array}{c} \alpha_* \\ \alpha_0 \\ \alpha_1 \end{array}\right] = \left[\begin{array}{c} y_{1_1} \\ \vdots \\ y_{\mu_1} \\ \hline y_{1_2} \\ \vdots \\ y_{\mu_2} \\ \hline y_{1_3} \\ \vdots \\ y_{\mu_3} \\ \hline \vdots \\ \hline y_{1_M} \\ \vdots \\ y_{\mu_M} \end{array}\right] \tag{9.3}$$

9.3 Solution

Once again the normal equations offer the easy path to solution as in (**??**). The first step is to compute the inverse of the product matrix. Recall that the dot product is a commutative operator; therefore only six of the nine matrix entries are unique:

$$\mathbf{A}^{\mathrm{T}}\mathbf{A} = \begin{bmatrix} \mathrm{J}\cdot\mathrm{J} & \mathrm{J}\cdot\mathbf{1} & \mathrm{J}\cdot x \\ \mathbf{1}\cdot\mathrm{J} & \mathbf{1}\cdot\mathbf{1} & \mathbf{1}\cdot x \\ x\cdot\mathrm{J} & x\cdot\mathbf{1} & x\cdot x \end{bmatrix} = \begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix}.$$

For clarity, the unique elements are specified:

$$\begin{array}{lll} a = \mathrm{J}\cdot\mathrm{J} & b = \mathrm{J}\cdot\mathbf{1} & c = \mathrm{J}\cdot x \\ & d = \mathbf{1}\cdot\mathrm{J} & e = \mathbf{1}\cdot x \\ & & f = x\cdot x \end{array}$$

In advance of the computing the inverse, first compute the determinant

$$\det\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right) = \Delta = 2bce + adf - ae^2 - c^2d - fb^2.$$

Using (**??**) the inverse is

$$\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1} = \Delta^{-1}\begin{bmatrix} df - e^2 & ce - bf & be - cd \\ \cdot & af - c^2 & bc - ae \\ \cdot & \cdot & ad - b^2 \end{bmatrix}.$$

The right-hand side in (9.3) is

$$\mathbf{A}^{\mathrm{T}} y = \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} \mathbf{J} \cdot y \\ \mathbf{1} \cdot y \\ x \cdot y \end{bmatrix}.$$

The least squares solution is provided as

$$\begin{bmatrix} \alpha_0 \\ \alpha_* \\ \alpha_1 \end{bmatrix} = \left(\mathbf{A}^{\mathrm{T}} \mathbf{A}\right)^{-1} \mathbf{A}^{\mathrm{T}} y$$

which distills down to

$$\alpha = \Delta^{-1} \begin{bmatrix} \beta_1 \left(df - e^2\right) + \beta_2 \left(ce - bf\right) + \beta_3 \left(be - cd\right) \\ \beta_1 \left(ce - bf\right) + \beta_2 \left(af - c^2\right) + \beta_3 \left(bc - ae\right) \\ \beta_1 \left(be - cd\right) + \beta_2 \left(bc - ae\right) + \beta_3 \left(ad - b^2\right) \end{bmatrix}.$$

The errors associate with the fit parameters are

$$\begin{bmatrix} \sigma_* \\ \sigma_0 \\ \sigma_1 \end{bmatrix} = \sqrt{\frac{r^{\mathrm{T}}}{(m-n)\,\Delta}} \sqrt{\begin{bmatrix} df - e^2 \\ af - c^2 \\ ad - b^2 \end{bmatrix}}.$$

The solutions are expressed in terms of dot products readily available in Fortran.

9.4 Problem Statement

9.5 Data

9.6 Results

9.6.1 Least Squares Results

9.6.2 Apex Angles

9.6.3 Qualitative Results

Table 9.2. *Problem statement for grain identification by rows (coupled linear regression).*

trial function	$y_{(\mu)}(x) = \mu\alpha_* + \alpha_0 + \alpha_1 x$	
merit function	$M(p) = \sum_{k=1}^{n} \left(y_k - \mu\alpha_* + \alpha_0 + \alpha_1 x_k\right)^2$	
number of zones	$m = 5$	
number of overlaps	$n = 4$	
rank defect	$m - n = 1$	
measurements	$\lambda = \{11, 13, 13, 13, 12\}$	
measurements	(x_k, y_k), $k = 1\colon 1024$	
system matrix	$\mathbf{A} = \begin{bmatrix} \mathbf{1} & x \end{bmatrix}$	
data vector	y	
linear system	(9.3)	
results	α_*	gap
	α_0	$y-$axis intercept
	α_1	slope
residual error	$r = \mathbf{A}\alpha - y$	

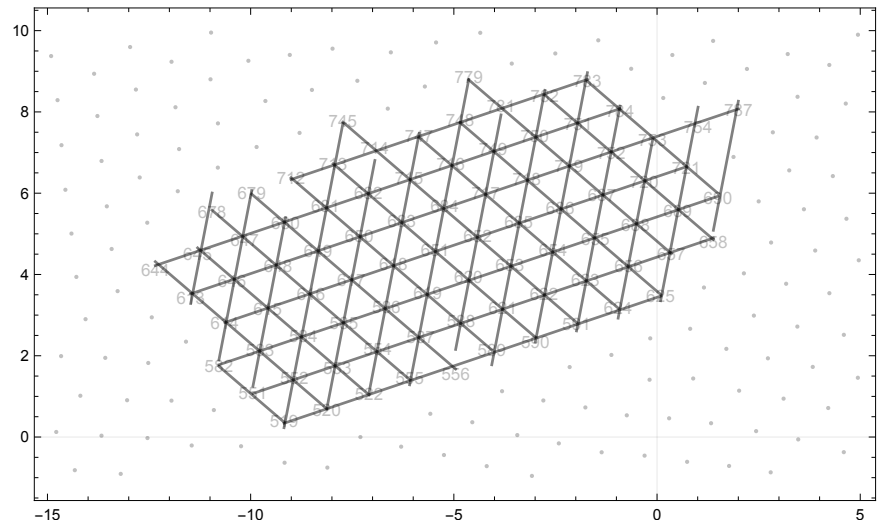


Figure 9.5. *Solutions for three data sets.*

Table 9.3. *Point membership in data sets shown in figure 9.1.*

set	row	1	2	3	4	5	6	7	8	9	10	11	12
1	1	519	520	522	555	556	589	590	591	624	625		
1	2	551	552	553	554	587	588	621	622	623	656	657	658
1	3	582	583	584	585	586	619	620	653	654	655	688	689
1	4	614	615	616	617	618	651	652	685	686	687	720	721
1	5	613	646	648	649	650	683	684	717	718	719	752	753
1	6	644	645	647	680	681	682	682	715	716	749	750	751
1	7	712	713	714	747	748	781	782	783				
2	1	658	690	787									
2	2	625	657	689	721	754							
2	3	624	656	688	720	753							
2	4	591	623	655	687	752	784						
2	5	590	622	654	686	719	751	783					
2	6	589	621	653	685	718	750	782					
2	7	556	588	620	652	717	749	781					
2	8	555	587	619	651	684	716	748	779				
2	9	522	554	586	618	683	715	747					
2	10	520	553	585	617	650	682	714					
2	11	519	552	584	616	649	681	713	745				
2	12	551	583	615	648	680							
2	13	582	614	646	647	679							
2	14	613	645	678									
3	1	582	551	519									
3	2	644	613	614	583	552	520						
3	3	645	646	615	584	553	522						
3	4	647	648	616	585	554	555	678					
3	5	679	680	649	617	586	587	556					
3	6	712	681	650	618	619	588	589					
3	7	713	682	683	651	620	621	590					
3	8	745	714	715	684	652	653	622	591				
3	9	747	716	717	685	654	623	624					
3	10	748	749	718	686	655	656	625					
3	11	779	781	750	719	687	688	657					
3	12	782	751	752	720	689	658						
3	13	783	784	753	721	690							

Table 9.4. *Excerpted data set.*

k	x_k	y_k	ϕ_k
1	−15.879001	−16.365496	−2.597531
2	−14.749446	−15.995488	−2.613017
3	−13.905339	16.242941	−2.557543
⋮			
1 022	13.927362	−16.235010	−2.780323
1 023	14.741765	15.957687	−2.687929
1 024	15.905518	16.346979	−2.599001

Table 9.5. *Least squares results for three axes.*

axis	gap			intercept			slope			$\sqrt{\langle r^2 \rangle}$
1	0.9899	±	0.0032	3.438	±	0.013	0.3376	±	0.0017	0.052
2	4.974	±	0.052	−2.075	±	0.093	5.168	±	0.052	0.18
3	1.2322	±	0.0039	−7.505	±	0.043	−0.8576	±	0.0038	0.054

Table 9.6. *Intermediate results: angles for the axes.*

axis	θ	±	σ_θ				
1	0.3256	±	0.0015	=	(18.655	±	0.086)°
2	1.380	±	0.018	=	(79.0	±	1.0)°
3	−0.7089	±	0.0025	=	(−40.62	±	0.14)°

Table 9.7. *Final results: apex angle measurements*

	θ	±	σ_θ				
α	1.040	±	0.018	=	(59.6	±	1.0)°
β	1.0345	±	0.0029	=	(59.27	±	0.17)°
γ	1.041	±	0.018	=	(59.7	±	1.0)°
total	3.116	±	0.026	=	(178.5	±	1.7)°

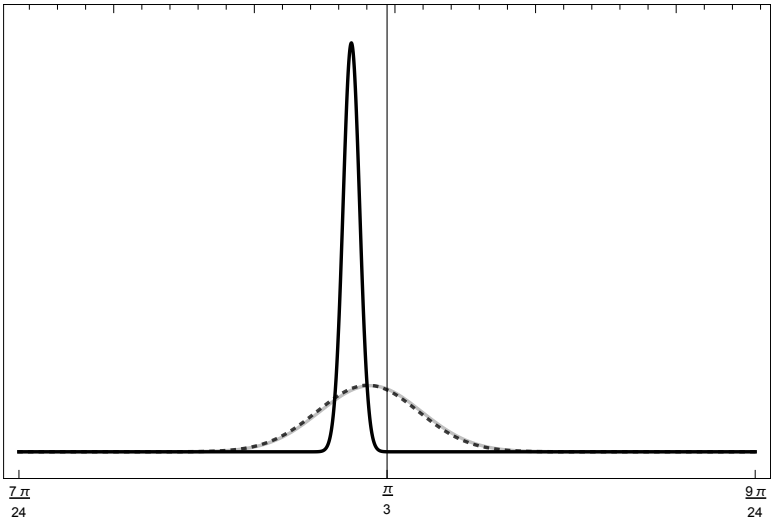


Figure 9.6. *Apex angles displayed in table 9.7.*

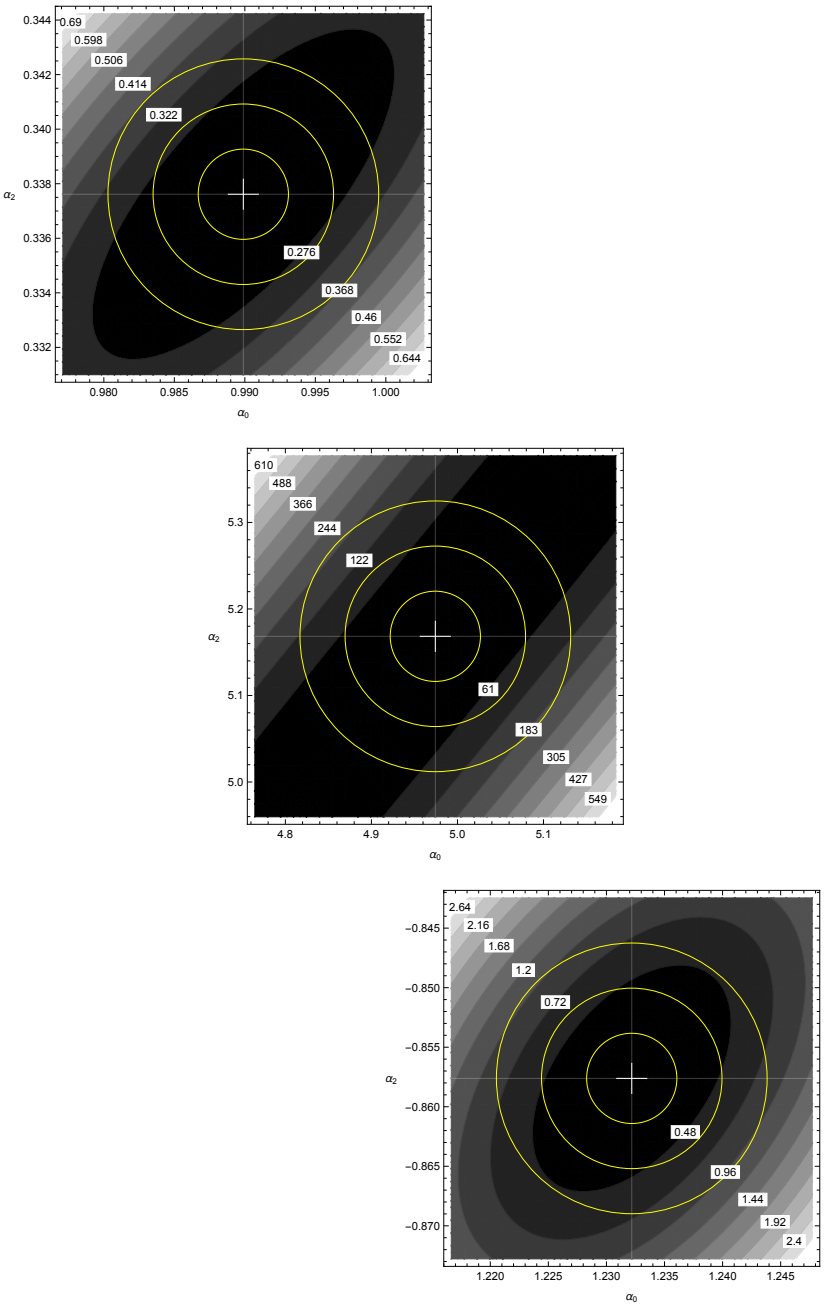


Figure 9.7. *Merit functions for the three data sets.*

Chapter 10

Crystals

In the previous model, the rows of atoms were treated independently. In this section the basic unit is not a row, it is instead a crystal. Mathematically, the process will imitate Nature: a seed crystal is picked, and other crystals will be identified from that.

Chapter 11

Stitching Local Maps

11.1 What is stitching?

Stitching is the process of combining local maps to create a global map.

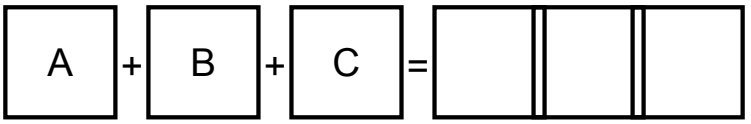


Figure 11.1. *Stitching local maps together to form a global map.*

1. ϕ
2. $\nabla\phi$
3. ϕ and $\nabla\phi$

11.2 Stitch ϕ

11.2.1 Genesis

$$\phi(x) = \exp\left(-\frac{x}{5}\right)\sin(\pi x)$$

11.2.2 Data

The central idea is simple; the mathematical expression is a tedious exercise in index gymnastics.

$$\zeta = 3$$

Table 11.1. *The input data in continuous and discrete form.*

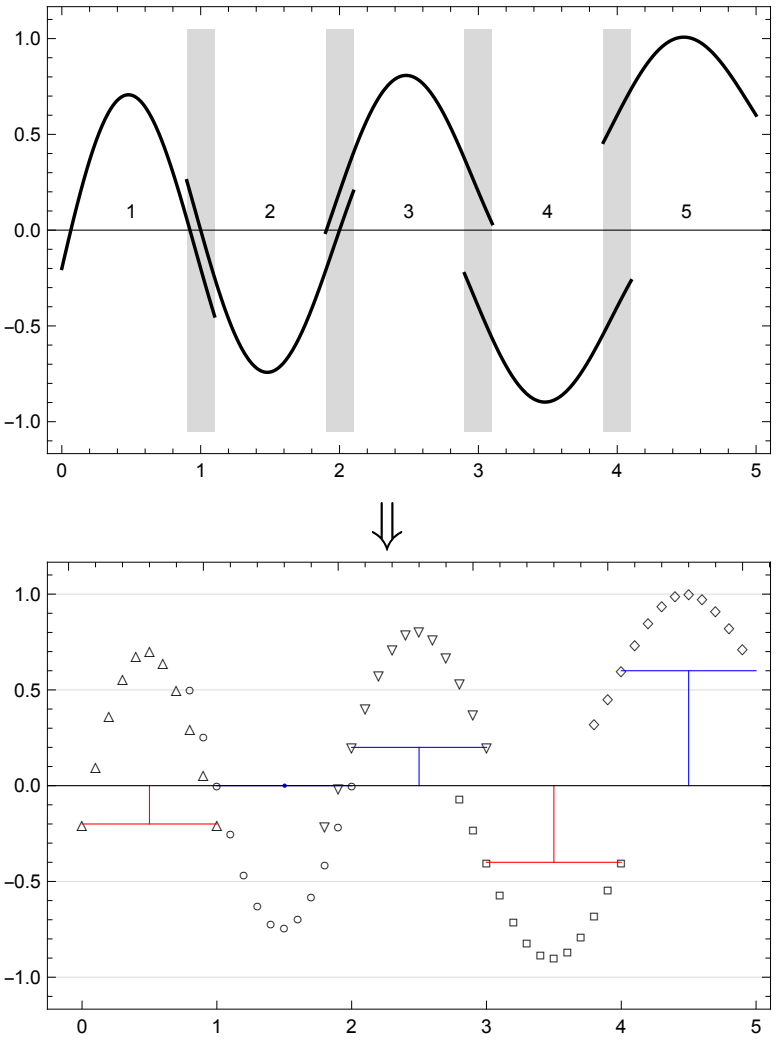


Table 11.2. *Sample showing an overlap of $\zeta = 3$ between the first two zones.*

$\phi_{1,1}$	$\phi_{2,1}$	$\cdots$	$\phi_{\lambda_1-2,1}$	$\phi_{\lambda_1-1,1}$	$\phi_{\lambda_1,1}$					
			$\phi_{1,2}$	$\phi_{2,2}$	$\phi_{3,2}$	$\cdots$	$\phi_{\lambda_1-2,2}$	$\phi_{\lambda_2-1,2}$	$\phi_{\lambda_2,2}$	

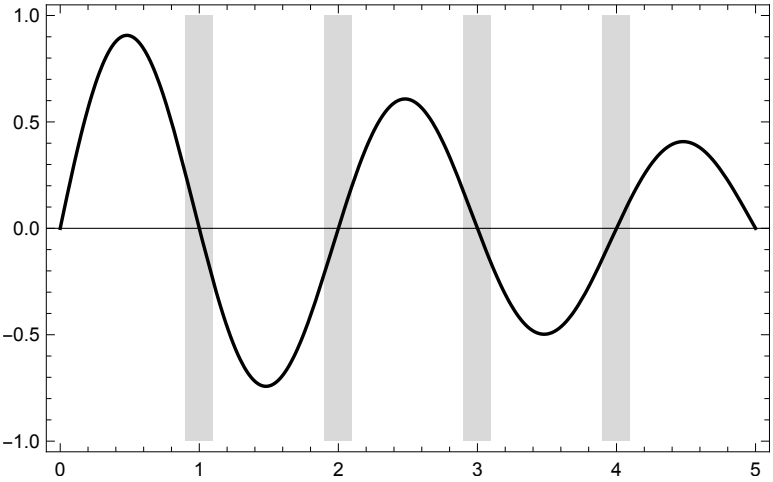


Figure 11.2. *The ideal potential function showing five measurement zones and four overlap bands.*

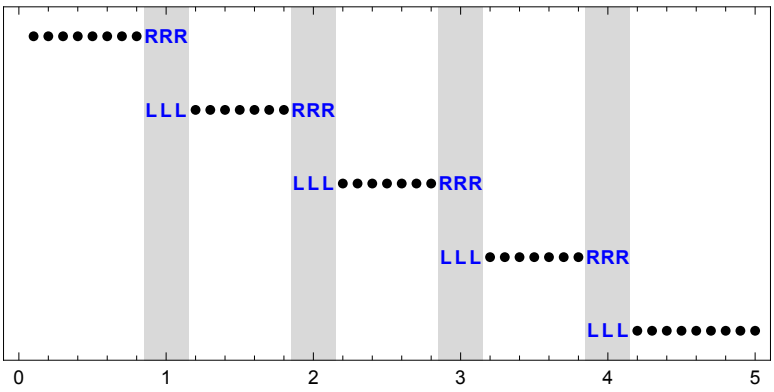


Figure 11.3. *Waterfall diagram showing discretization within measurement zones with left and right zone overlaps.*

$$\Delta_{12} = \zeta^{-1}\left(\underbrace{(\phi_{\lambda_1-2,1} + \phi_{\lambda_1-1,1} + \phi_{\lambda_1,1})}_{\text{zone 1}} - \underbrace{(\phi_{1,2} + \phi_{2,2} + \phi_{3,2})}_{\text{zone 2}}\right)$$

$$\Delta_{12} = \zeta^{-1}\left(\underbrace{(\phi_{\lambda_1-2,1} - \phi_{1,2})}_{\text{pair 1}} + \underbrace{(\phi_{\lambda_1-1,1} - \phi_{2,2})}_{\text{pair 2}} + \underbrace{(\phi_{\lambda_1,1} - \phi_{3,2})}_{\text{pair 3}}\right)$$

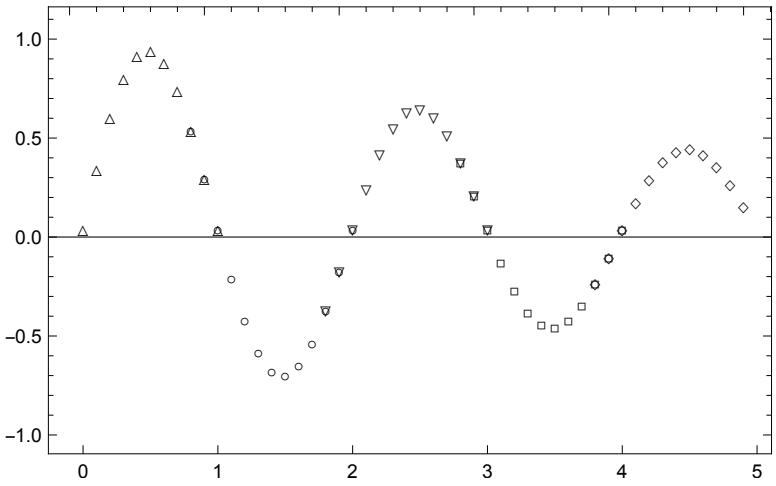


Figure 11.4. *Stitching unifies the data.*

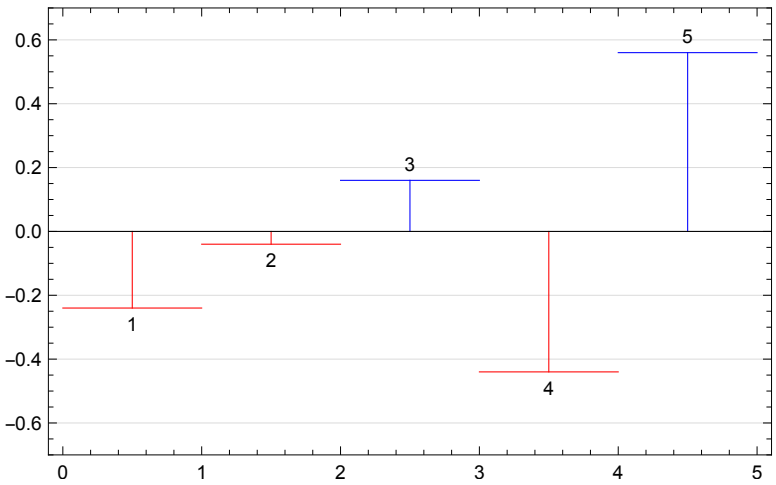


Figure 11.5. *A set of piston adjustments which restores continuity across the domain.*

Mean value of the differences.

$$\Delta_{j,j+1} = \zeta^{-1} \sum_{k=1}^{\zeta} p_{j,\lambda_j - \zeta + k} - p_{j+1,k}$$

k	y_1	y_2	y_3	y_4	y_5
1	−0.2	0.500 878	−0.210 084	−0.064 251 7	0.325 113
2	0.102 898	0.258 113	−0.011 324 8	−0.226 982	0.458 345
3	0.364 738	0.0	0.2	−0.4	0.6
4	0.561 904	−0.247 992	0.403 039	−0.566 234	0.736 101
5	0.677 936	−0.462 368	0.578 555	−0.709 935	0.853 753
6	0.704 837	−0.623 794	0.710 719	−0.818 142	0.942 345
7	0.643 511	−0.718 793	0.788 498	−0.881 821	0.994 482
8	0.503 326	−0.740 818	0.806 531	−0.896 585	1.006 57
9	0.300 878	−0.690 609	0.765 423	−0.862 929	0.979 014
10	0.058 112 7	−0.575 834	0.671 453	−0.785 993	0.916 025
11	−0.2	−0.410 084	0.535 748	−0.674 887	0.825 059
12		−0.211 325	0.373 018	−0.541 655	0.715 978
13		0.0	0.2	−0.4	

Table 11.3. *Measurements displaying the connection between overlap bands in figure 11.3.*

First overlap region.

$$\Delta_{12} = \zeta^{-1}\left(\underbrace{(\phi_{9,1} + \phi_{10,1} + \phi_{11,1})}_{\text{last 3 elements of zone 1}} - \underbrace{(\phi_{1,2} + \phi_{2,2} + \phi_{3,2})}_{\text{first 3 elements of zone 2}}\right)$$

11.2.3 Data and results

Table 11.4. *Computation of the zone shift values.*

$$\begin{aligned}
\Delta_{12} &= \tfrac{1}{3}\left((\phi_{9,1} + \phi_{10,1} + \phi_{11,1}) - (\phi_{1,2} + \phi_{2,2} + \phi_{3,2})\right) \\
\Delta_{23} &= \tfrac{1}{3}\left((\phi_{11,2} + \phi_{12,2} + \phi_{13,2}) - (\phi_{1,3} + \phi_{2,3} + \phi_{3,3})\right) \\
\Delta_{34} &= \tfrac{1}{3}\left((\phi_{11,3} + \phi_{12,3} + \phi_{13,3}) - (\phi_{1,4} + \phi_{2,4} + \phi_{3,4})\right) \\
\Delta_{45} &= \tfrac{1}{3}\left((\phi_{11,4} + \phi_{12,4} + \phi_{13,4}) - (\phi_{1,5} + \phi_{2,5} + \phi_{3,5})\right)
\end{aligned}$$

11.2.4 Linear System

$$\mathbf{A}p = \Delta$$

Table 11.5. *Computation of the zone shift values.*

$$
\begin{aligned}
\Delta_{12} &= \tfrac{1}{3}\left((0.300878 + 0.0581127 - 0.2) - (0.500878 + 0.258113 + 0.)\right) \\
\Delta_{23} &= \tfrac{1}{3}\left((-0.410084 - 0.211325 + 0.) - (-0.210084 - 0.0113248 + 0.2)\right) \\
\Delta_{34} &= \tfrac{1}{3}\left((0.535748 + 0.373018 + 0.2) - (-0.0642517 - 0.226982 - 0.4)\right) \\
\Delta_{45} &= \tfrac{1}{3}\left((-0.674887 - 0.541655 - 0.4) - (0.325113 + 0.458345 + 0.6)\right)
\end{aligned}
$$

Table 11.6. *Input data*

	Shift	Value
1	Δ_{12}	–0.2
2	Δ_{23}	0.0
3	Δ_{34}	0.6
4	Δ_{45}	–1.

$$
\begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{bmatrix} = \begin{bmatrix} \Delta_{12} \\ \Delta_{23} \\ \Delta_{34} \\ \Delta_{45} \end{bmatrix}
$$

$$
p_{LS} = \frac{1}{5} \begin{bmatrix} 4 & 3 & 2 & 1 \\ -1 & 3 & 2 & 1 \\ -1 & -2 & 2 & 1 \\ -1 & -2 & -3 & 1 \\ -1 & -2 & -3 & -4 \end{bmatrix} \begin{bmatrix} \Delta_{12} \\ \Delta_{23} \\ \Delta_{34} \\ \Delta_{45} \end{bmatrix} + \alpha \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}
$$

$$
\mathbf{A}^{\dagger} b = \frac{1}{25} \begin{bmatrix} -6 \\ -1 \\ 4 \\ -11 \\ 14 \end{bmatrix}
$$

These are the actual plot values used in figure 11.5.

$$
\Phi_{corrected} = \Phi_{measured} - \mathbf{A}^{\dagger} b
$$

Table 11.7. *Problem statement for linear regression.*

trial function	$p_k - p_{k+1} = \Delta_{k,k+1},\ k = 1\colon n$
merit function	$M(p) = \sum_{k=1}^{n} \left(\Delta_{k,k+1} - p_k + p_{k+1}\right)^2$
number of zones	$m = 5$
number of overlaps	$n = 4$
rank defect	$m - n = 1$
measurements per zone	$\lambda = \{11, 13, 13, 13, 12\}$
measurements	$\phi_{k,j},\ k = 1\colon m,\ j = 1\colon \lambda_m$
input data	$\Delta_{k,k+1},\ k = 1\colon n$
results	$p_k,\ k = 1\colon m$
residual error	$r = \mathbf{A}^{\dagger} b - \Delta$
linear system	$\begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{bmatrix} = \begin{bmatrix} \Delta_{12} \\ \Delta_{23} \\ \Delta_{34} \\ \Delta_{45} \end{bmatrix}$
gauge condition	$\sum_{k=1}^{m} p_k = 0$

11.2.5 Least Squares Arbitration

There is a fundamental ambiguity arising from gradient measurements stemming from the basic fact that

$$\frac{d}{dx}\phi(x) = \frac{d}{dx}\left(\phi(x) + c\right).$$

We can recover the function shape, but not the offset. In other words, there is a translation invariance. This lone constant is the poster child for the rank one deficiency in the linear system of (11.2.4). Realizing this, the one dimensional problem could be solved without resort to least squares.

The system can be solved, for example, by moving from left to right and manually forcing the data to match. If the the overlap difference between zone 1 and zone 2 is Δ_{12}, add Δ_{12} to every value in zone 2. Now zones 1 and 2 are stitched together. Compute Δ_{23}, add this value to every point in zone 3. Zones 1, 2, and 3 are now stitched together. Continue as needed.

The least squares problem is obviated. How did this happen? The process of least squares is an exercise error arbitration which takes a peanut butter approach by trying to distribute the error evenly. In one dimension, there is no need for arbitration as there is no conflict in measurements.

In two dimensions, the problem changes. Consider the typical cell with a neighbor to the right and a neighbor above. The right–left overlap adjustment conflicts with the up–down overlap adjustment. The least squares process takes

Table 11.8. *Results for stitching with piston.*

fit parameters	p_k, $k = 1\colon m$	pistons
computed solution	$p = \frac{1}{25}\begin{bmatrix} -6 \\ -1 \\ 4 \\ -11 \\ 14 \end{bmatrix}$	$\mathbf{A}^\dagger b$
data vector	$\frac{1}{5}\begin{bmatrix} -1 \\ 0 \\ 1 \\ -2 \\ 3 \end{bmatrix} = \frac{1}{25}\begin{bmatrix} -6 \\ -1 \\ 4 \\ -11 \\ 14 \end{bmatrix} - \frac{1}{5}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$	$p_{\mathcal{R}} + p_{\mathcal{N}}$
residual error	$r \cdot r = 0$	
problem statement	table 11.7	
measurements	table 11.3	
input data	table 11.6	
plots	figure 11.2	raw data (bottom)
	figure 11.4	corrected data

all off the overlap conflicts and provides a set of adjustments which minimizes the global error. To close, note that the least squares solution was used even though it is not necessary until dimension 2 or higher.

One last tidbit. Figure 11.7 shows the piston values that were input to distort the values. Least squares chooses a distinct set of corrections. Why was this set selected? A tantalizing clue is given by the null space vector in (11.2.4). Notice this vector is perpendicular to every column vector in $\mathbf{A}^\dagger$ which implies that the sum of each column vector must be 0. Therefore, the gauge condition is that the solution vector will have sum 0:

$$p_1 + p_2 + p_3 + p_4 + p_5 = 0.$$

We may now eliminate a variable; choose the last one:

$$p_5 = -p_1 - p_2 - p_3 - p_4$$

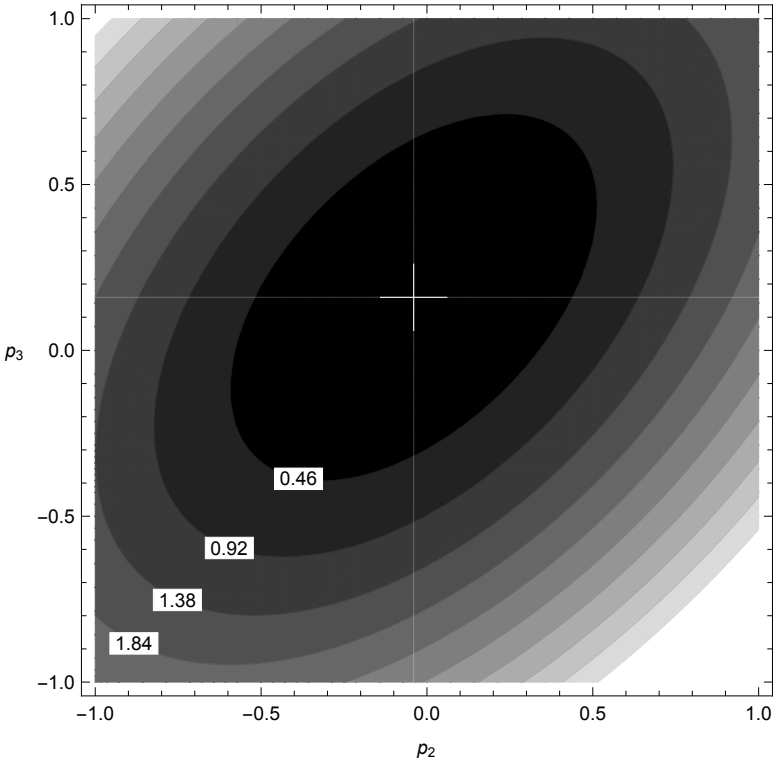


Figure 11.6. *Looking at the merit function on the $p_2 - p_3$ axis.*

Instead of (11.2.4), there is now

$$\begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ 1 & 1 & 1 & 2 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix} = \begin{bmatrix} \Delta_{12} \\ \Delta_{23} \\ \Delta_{34} \\ \Delta_{45} \end{bmatrix}.$$

The solution is the same:

$$\hat{p}_{gauge} = \begin{bmatrix} 4 & 3 & 2 & 1 \\ -1 & 3 & 2 & 1 \\ -1 & -2 & 2 & 1 \\ -1 & -2 & -3 & 1 \end{bmatrix} \begin{bmatrix} \Delta_{12} \\ \Delta_{23} \\ \Delta_{34} \\ \Delta_{45} \end{bmatrix}.$$

The 0 sum, or equivalently 0 mean, condition is a gauge condition which restores the column rank of the problem.

The piston values used to create the data set are decomposed into range and

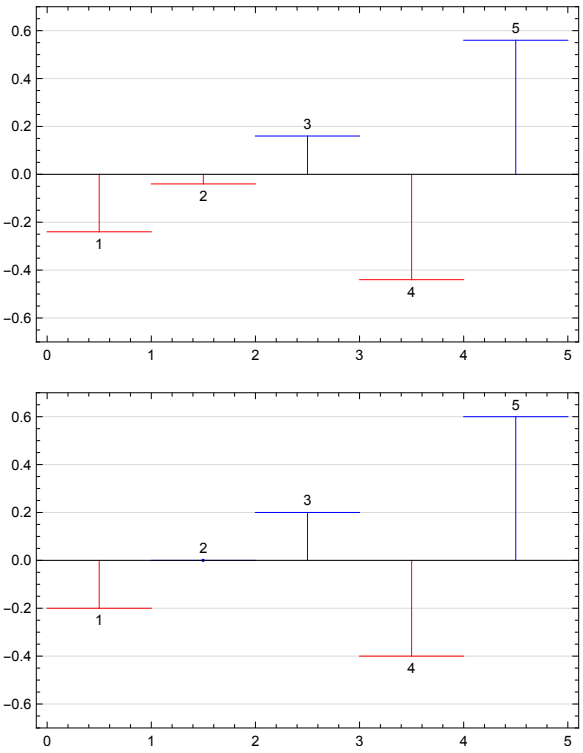


Figure 11.7. *On top, pistons output from the solution; on bottom, pistons input to create the data.*

null space terms.

$$\frac{1}{5}\begin{bmatrix} -1 \\ 0 \\ 1 \\ -2 \\ 3 \end{bmatrix} = \frac{1}{25}\begin{bmatrix} -6 \\ -1 \\ 4 \\ -11 \\ 14 \end{bmatrix} - \frac{1}{5}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

11.3 Stitch $\nabla\phi$

The next challenge is to stitch data together using the gradient $\nabla\phi$ rather than the function value ϕ. The outputs now will be a set of piston adjustments called tilts which restore continuity of the gradient.

The problem arises in the field of wavefront sensing. Modern devices make exquisite measurements of tilts. The process of wavefront reconstruction takes these tilts and reconstructs the wavefront. Measure $\nabla\phi(x)$ and compute $\nabla\phi(x)$.

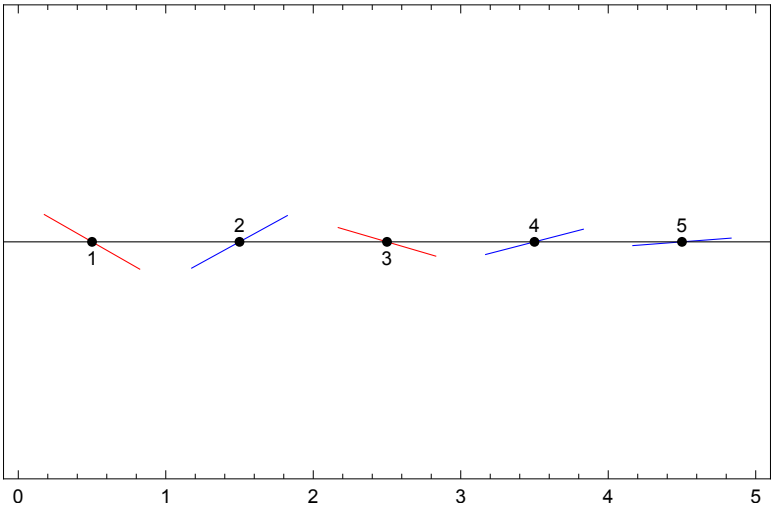


Figure 11.8. *A set of tilt adjustments which restores continuity of the gradient across the domain.*

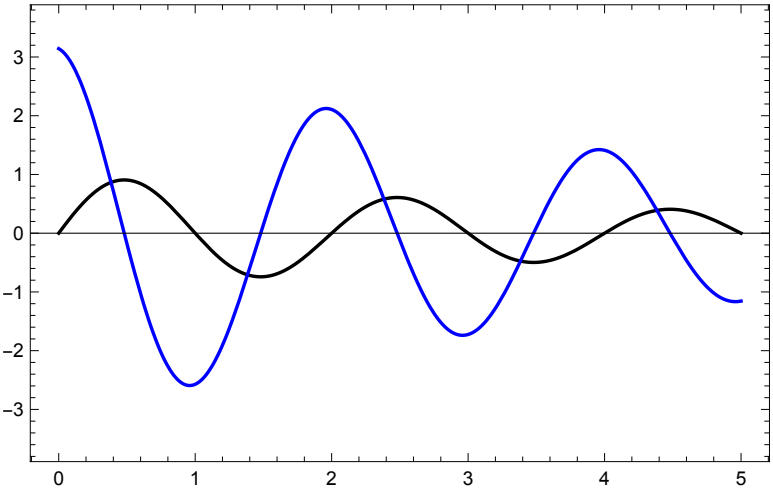


Figure 11.9. *A function (black) and its gradient (blue).*

Gradient of (11.3)

$$\nabla\phi(x) = \frac{1}{5}\exp\left(-\frac{x}{5}\right)(5\pi\cos(\pi z) - \sin(\pi z))$$

$$\tau = \frac{1}{100}\begin{bmatrix} -90 \\ 85 \\ -40 \\ 35 \\ 10 \end{bmatrix}$$

A scaled version of these values is plotted in figure 11.8.

Chapter 12

Gradient I

$$F = \nabla \phi$$

$$W^{1,2}(\Omega) = \left\{ \phi \in L^2(\Omega) : \partial_x^1 \phi \in L^2(\Omega) \right\}$$

12.1 One Dimension

$$\Omega = \bigcup_k \omega_k$$

Interval

$$\omega = \{ x \in \mathbb{R} : a < x < b \}$$

Average gradient

$$\langle \nabla \phi(x) \rangle_\omega = \phi(b) - \phi(a)$$

$$\begin{bmatrix} -1 & 1 & 0 & \dots & & \\ 0 & -1 & 1 & & & \\ \vdots & & \ddots & \ddots & & \vdots \\ & & & -1 & 1 & 0 \\ & & & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} \varphi_0 \\ \varphi_1 \\ \vdots \\ \varphi_{m-1} \\ \varphi_m \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}$$

Part IV

Applications: Nonlinear Problems

Can I write something here?

Chapter 13

Finding the Best Circle

Find the circle from edge points.

Table 13.1. *Problem statement for linear regression.*

trial function	$\rho^2 = (p-O)^T (p-O)$	L^2
residual error	$r_k = \rho^2 - (p_k - O)^T (p_k - O)$	L^2
merit function	$M(a) = \sum_{k=1}^{m} \left(\rho^2 - (p_k - O)^T (p_k - O)\right)^2$	$(^\circ\mathrm{C})^2$
measurements	p_k, $k = 1{:}m$	position, $\mathbb{R}^n$
results	$O \pm \epsilon_O$	origin, $\mathbb{R}^n$
	$\rho^2 \pm \epsilon_{\rho^2}$	radius, L
# 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 4.2	

Table 13.2. *Results for best circle*

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 15.3
	residuals: figure 15.4
	merit function in $\mathcal{R}(\mathbf{A}^*)$: figure 15.6

Chapter 14

Linearization

14.1 Powers Laws and Exponentials

Exponential relationships, such as radioactive decay,

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

and power laws, like the learning curve,

$$y = ax^b \tag{14.1}$$

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.

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

14.2 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

$$M(a, b) = \sum_{k=1}^{m} \left(y_k - ak^b \right)^2 \tag{14.2}$$

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 14.2. What does the surface of the merit function in (14.2) 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.

14.2.1 Problem Statement

The problem stated in table 14.1 embodies the major points discussed up to this point and displays a few details that will become clear as the discussion advances.

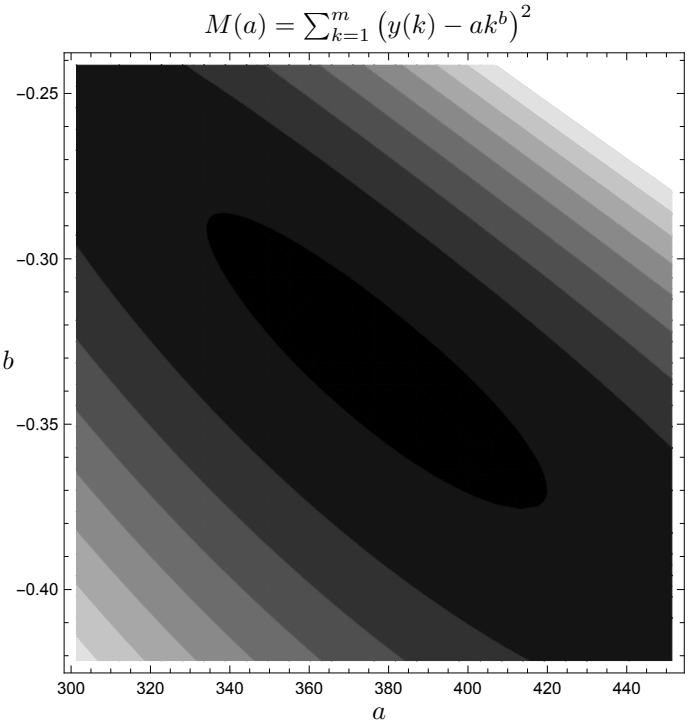


Figure 14.1. *Merit function for the learning curve in solution space.*

14.2.2 Solution

Without a linear system, attack with calculus as in §4.2.2. 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 14.2. 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 14.1. *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 14.2	

Table 14.2. *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:

$$\hat{M}(b) = \sum \left(y(x) - \frac{\sum y_k k^b}{\sum k^{2b}} k^b \right)^2 \tag{14.3}$$

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

Figure 14.2 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 14.1. The curve in figure 14.2 represents a slice through the surface of the merit function in figure 14.1. Such a slice is shown in figure 14.3.

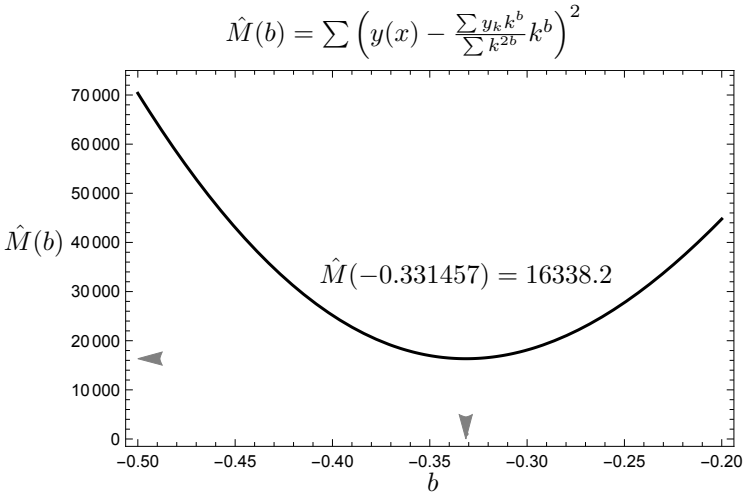


Figure 14.2. *Merit function constrained to one parameter b.*

14.2.3 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 14.2.

The merit function in figure 14.8 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.

14.3 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.

14.3.1 Logarithmic Transform

While the logarithmic transformation is exact,

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

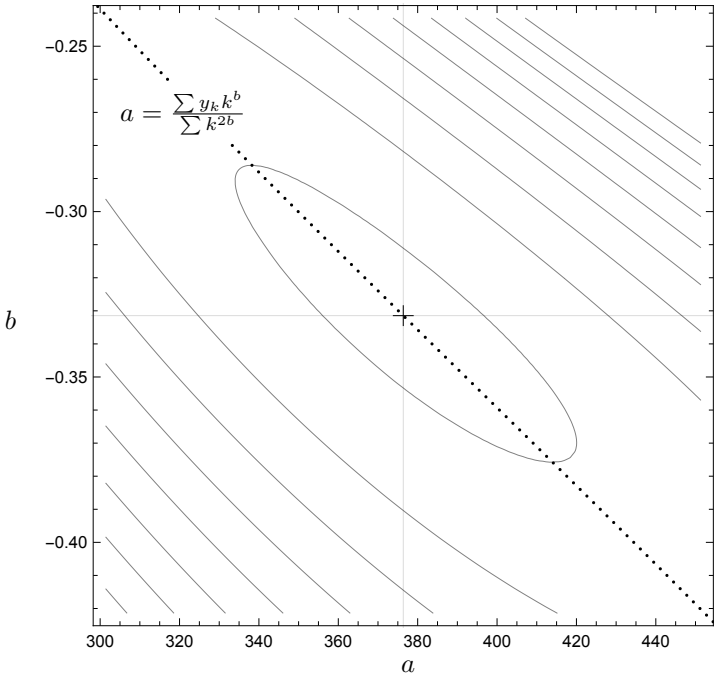


Figure 14.3. *Merit function for the learning curve in solution space showing the constrained a parameter as a dotted line.*

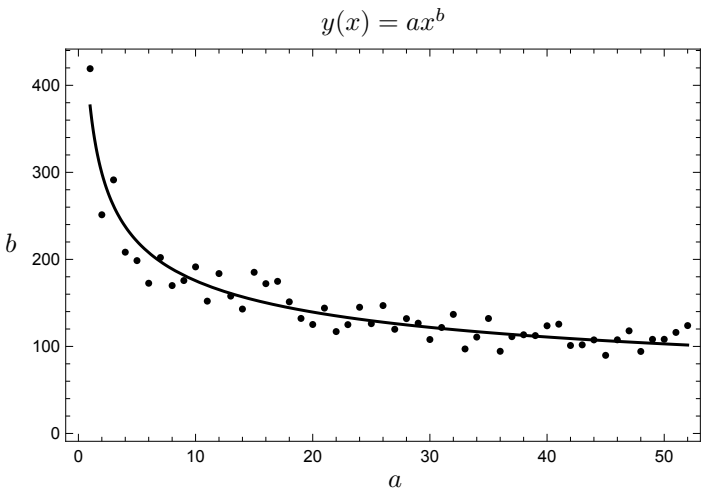


Figure 14.4. *Solution for equations* (14.2) *and* (14.3) *using data in table 14.2.*

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 14.3. *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 14.1	
input data	table 14.2	
plots	figure 14.4	1. data and solution
	figure 14.5	2. residual errors
	figure 14.8	3. merit function

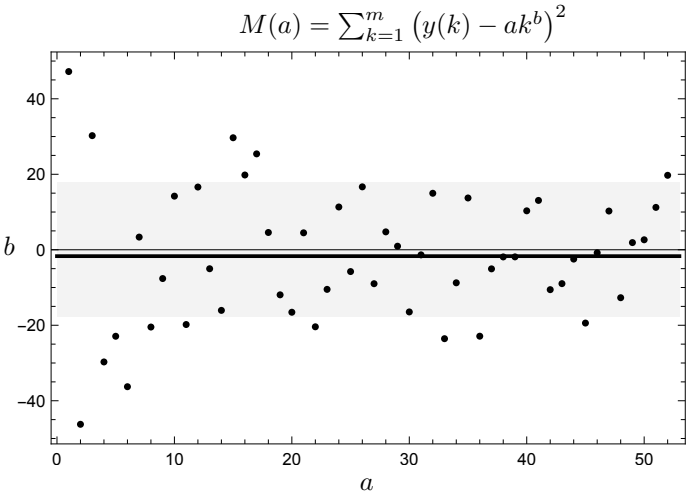


Figure 14.5. *The residual errors in figure 14.4.*

the ratio of two distinct data points,

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

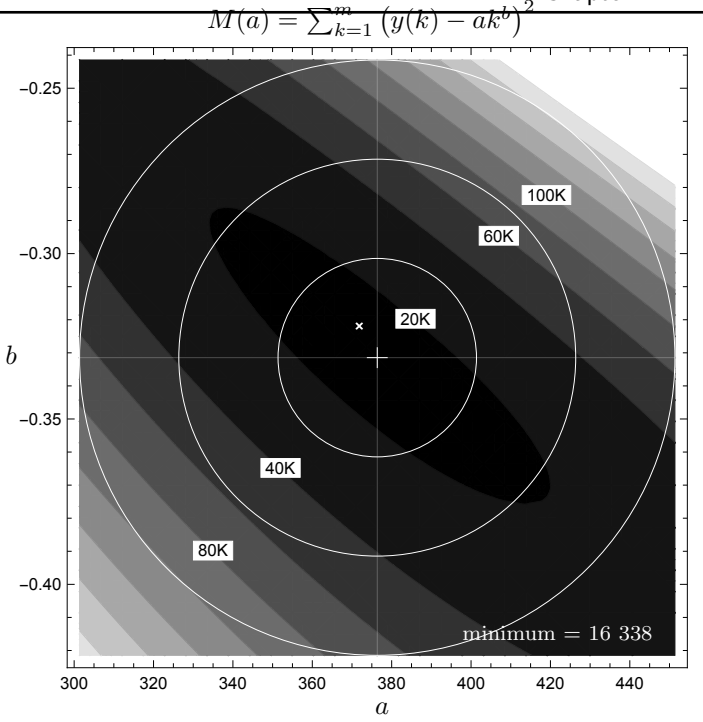


Figure 14.6. *Minimization of the merit function for the learning curve.*

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.

14.3.2 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

$$\begin{aligned} \ln a = 5.88094 \qquad &\rightarrow \qquad a = 358.146, \\ b = -0.314929. \end{aligned} \tag{14.4}$$

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

$$M(358.146, -0.314929) = 17005 > 16338. \tag{14.5}$$

The eyeball is a poor instrument for distinguishing between the correct solution in table 14.6 and (14.4). However, the merit function sharply delineates the two.

The merit shows the solution (14.4) marked with ⊞, distinctly away from the minimum of the merit function as seen in equation (14.5) and figure 14.8.

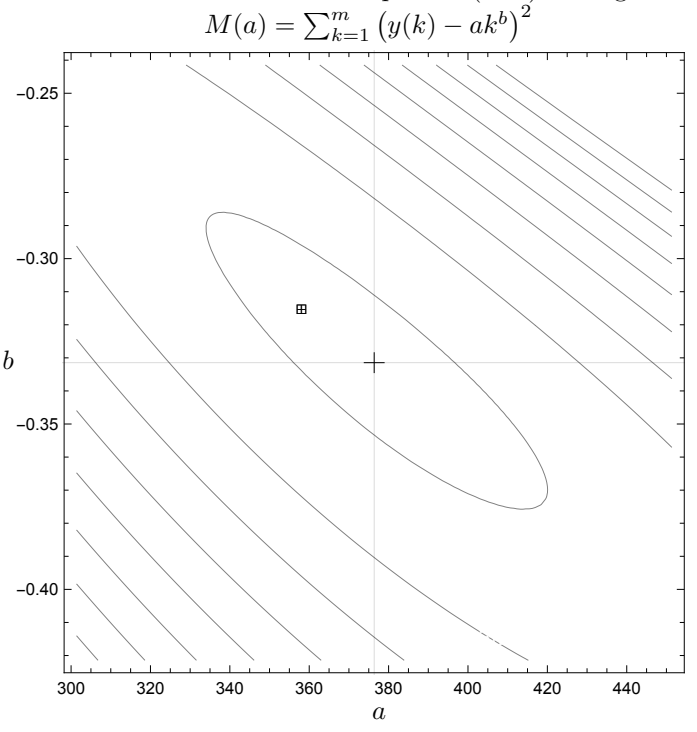


Figure 14.7. *The merit function for the learning curve showing the minimum and the value .*

14.3.3 Reflection Test Fails

14.4 Radioactive Decay

14.4.1 Theory

14.4.2 Problem Statement

Table 14.4. *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 4.2	

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

14.4.3 Results

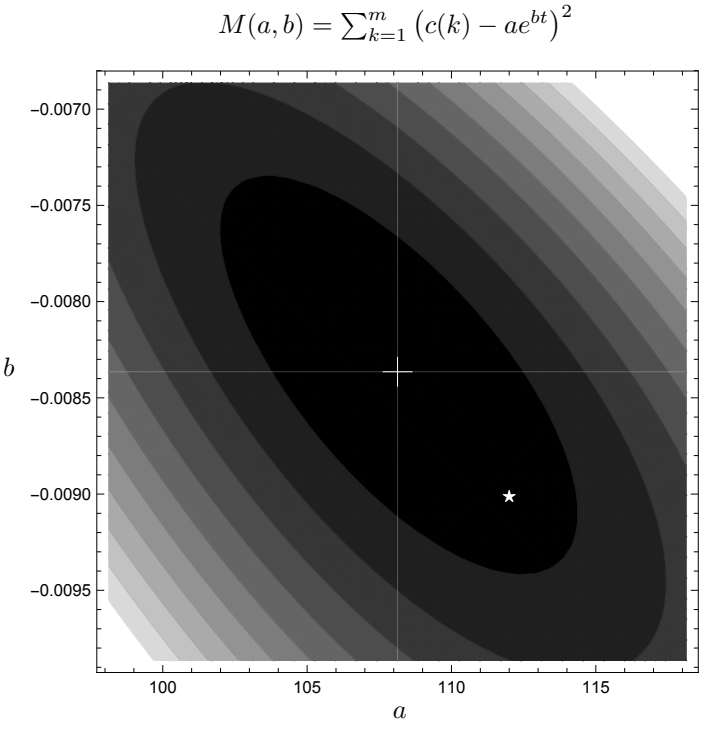
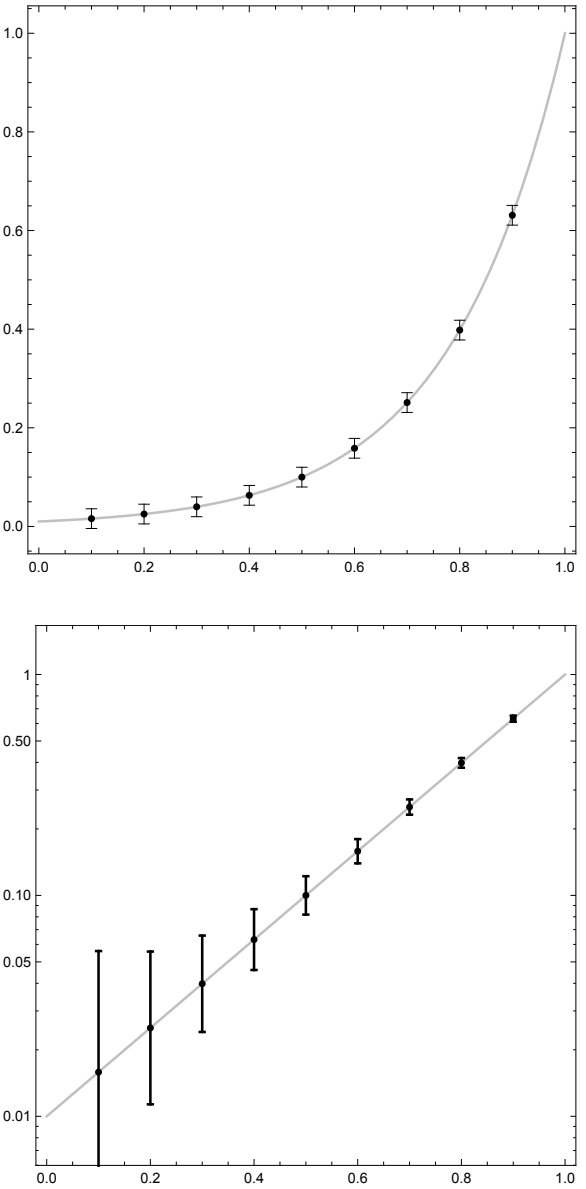


Figure 14.8. *The merit function for the radioactive decay model showing the minimum and the value .*

Table 14.5. *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 14.6. *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 14.4	
input data	table 4.2	
plots	figure 4.2	1. data and solution
	figure 4.3	2. residual errors
	figure **??**	3. merit function

Chapter 15

Population Growth

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

15.1 Model

$$y(\tau) = a_1 + a_2\tau + a_3 e^{d\tau} \tag{15.1}$$

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

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

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

$$\min_{\substack{a\in\mathbb{R}^3 \\ d\in\mathbb{R}}} \left\| \mathbf{A}(d) \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} - y \right\|_2^2 \tag{15.2}$$

Table 15.1. *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 15.2
results	a_0	constant
	a_1	linear
	a_2	exponential
	d	power term
residual error	$r = \mathbf{A}^{\dagger} b - \Delta$	
linear system	$\mathbf{A}(d) \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = y$	

15.2 Problem Statement

$$r^2 = \left\| \mathbf{A}(d) \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} - y \right\|_2^2 \tag{15.3}$$

15.3 Data

15.4 Example

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

15.5 Polynomials

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

Table 15.2. *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%

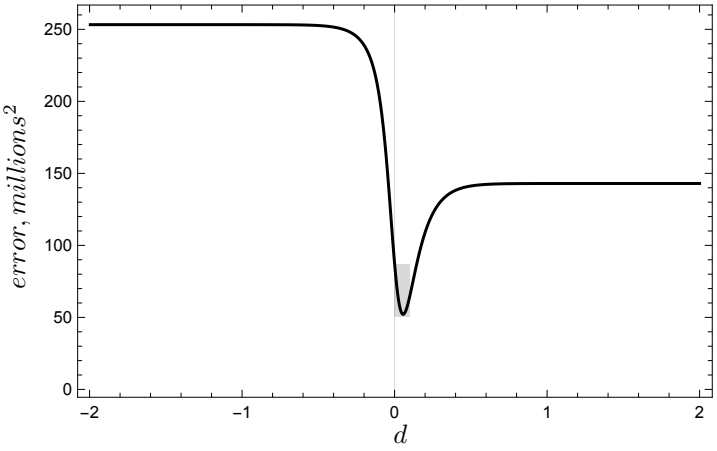


Figure 15.1. *Residual error for* (15.3) *with* a_1, a_2, *and* a_3 *at optimal values.*

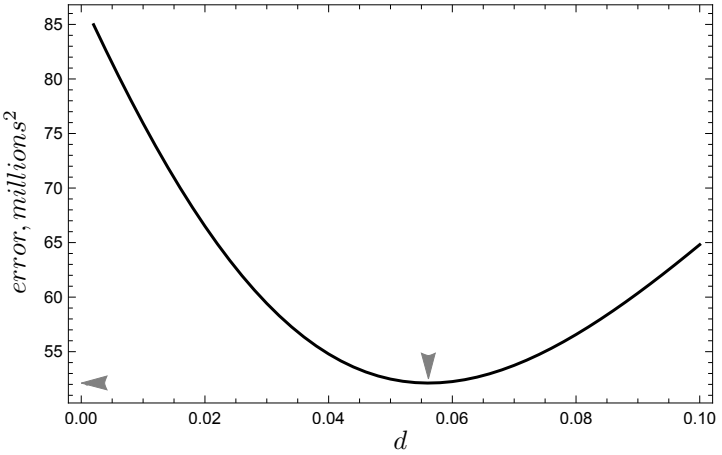


Figure 15.2. *Residual error for a_1 and a_2 fixed at optimal values.*

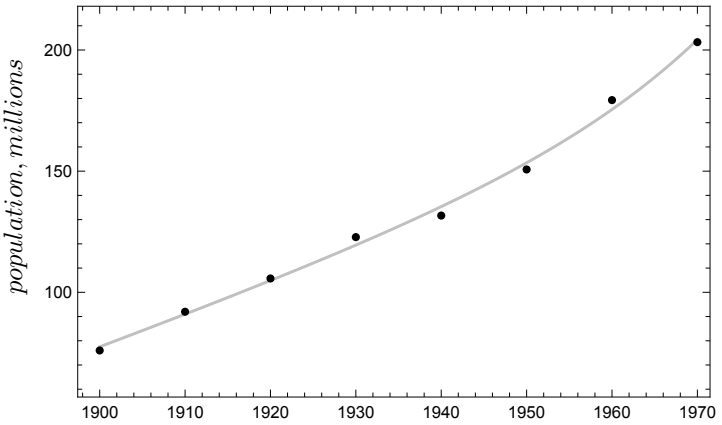


Figure 15.3. *Solution plotted against data.*

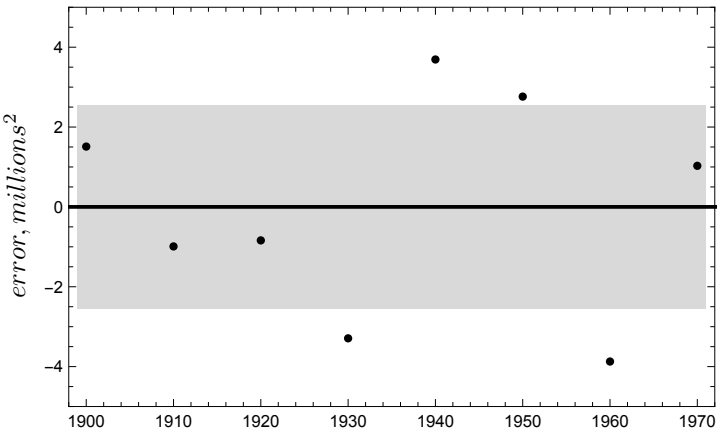


Figure 15.4. *Scatterplot of residual errors.*

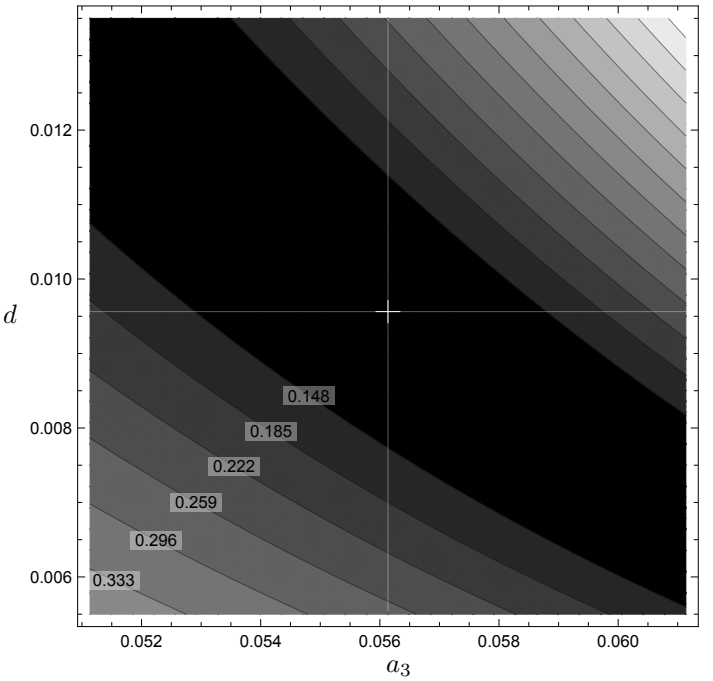


Figure 15.5. *The merit function with a_1 and a_2 fixed at best values showing least squares solution (center cross).*

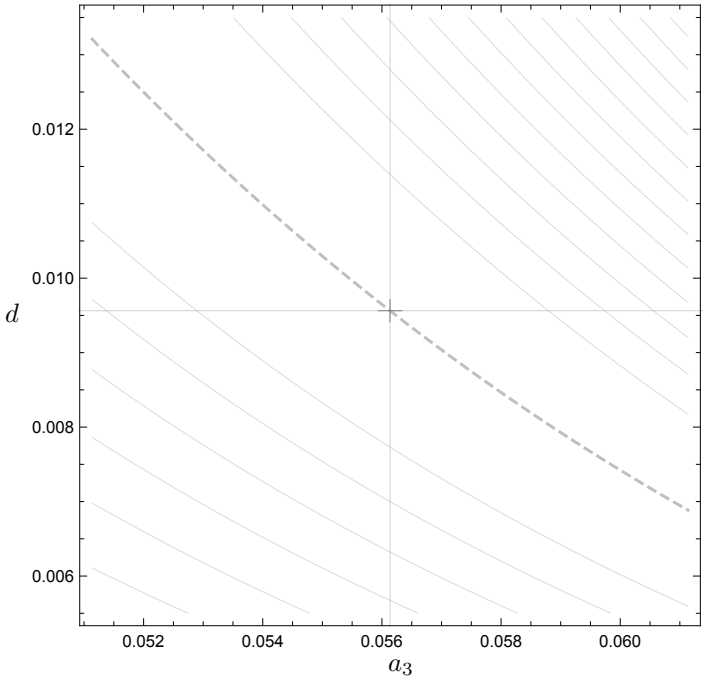


Figure 15.6. *The merit function with a_1 and a_2 fixed at best values showing least squares solution (center cross) and the (dashed line).*

Table 15.3. *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 15.3
	residuals: figure 15.4
	merit function in $\mathcal{R}(\mathbf{A}^*)$: figure 15.6

Table 15.4. *Fitting the census data with low order polynomials.*

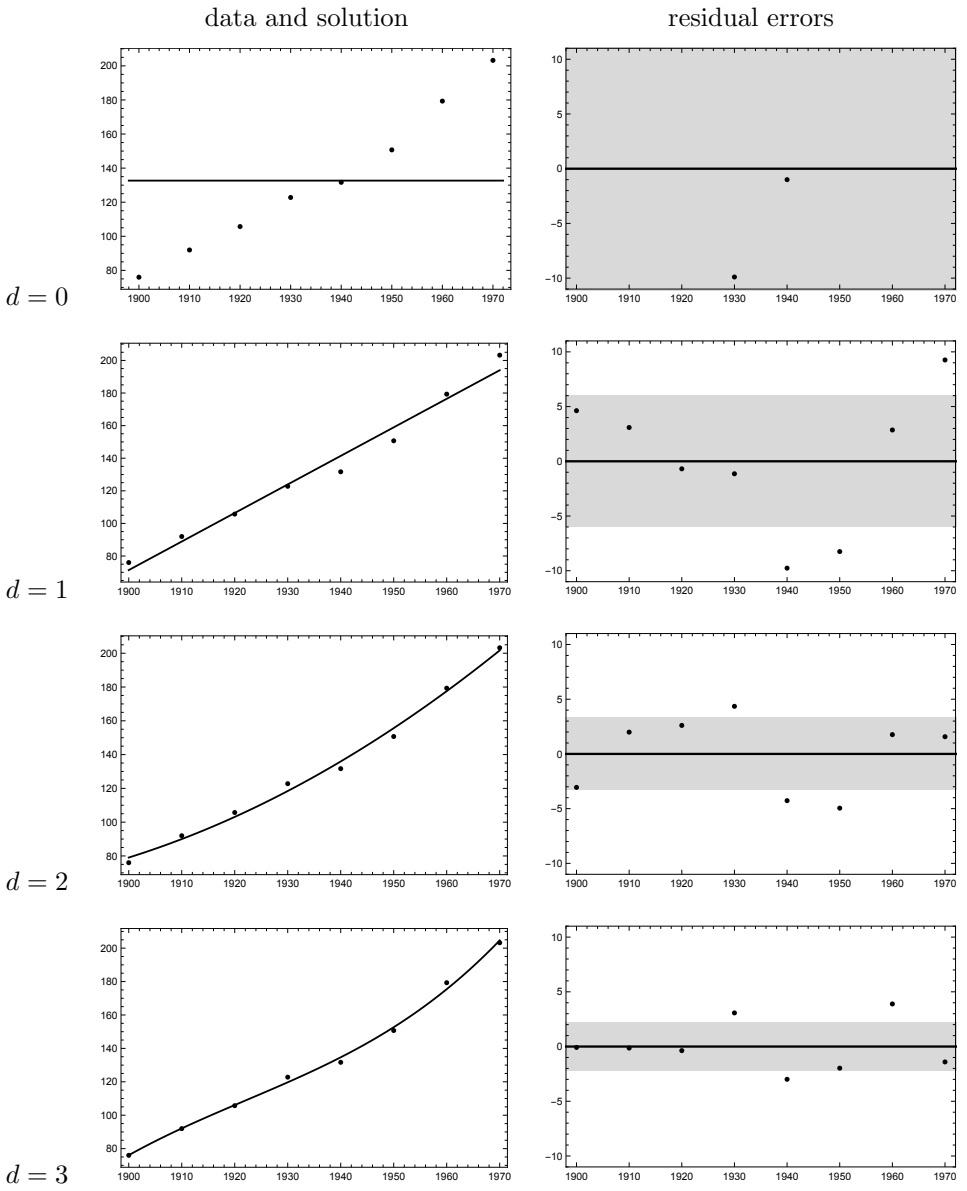


Table 15.5. *Fitting the census data with higher order polynomials.*

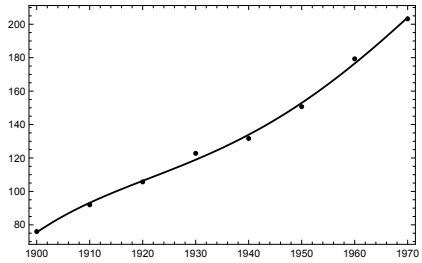
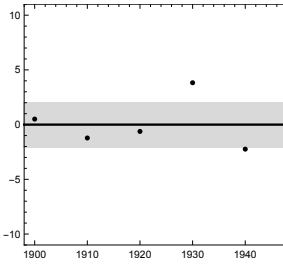
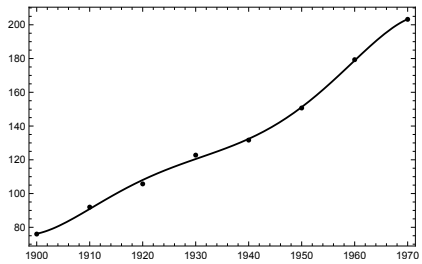
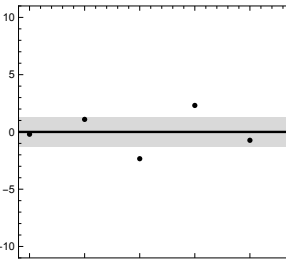
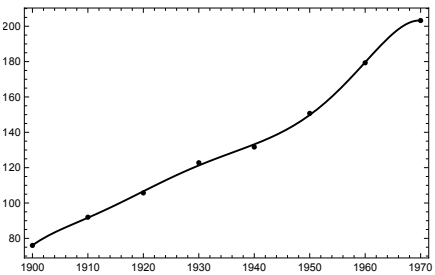
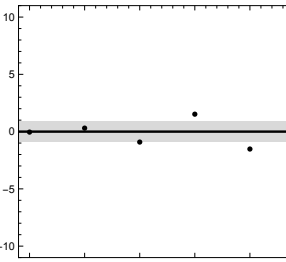
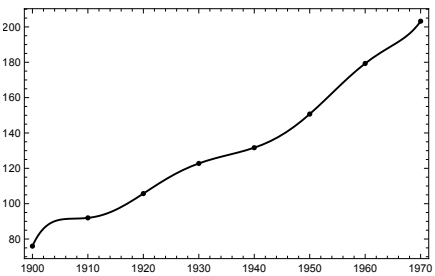
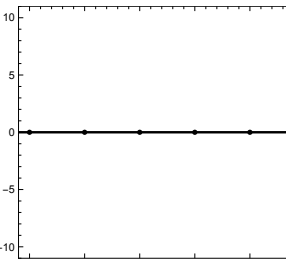
	data and solution	residual errors
$d = 4$		
$d = 5$		
$d = 6$		
$d = 7$		

Table 15.6. *Fitting the census data with low order polynomials.*

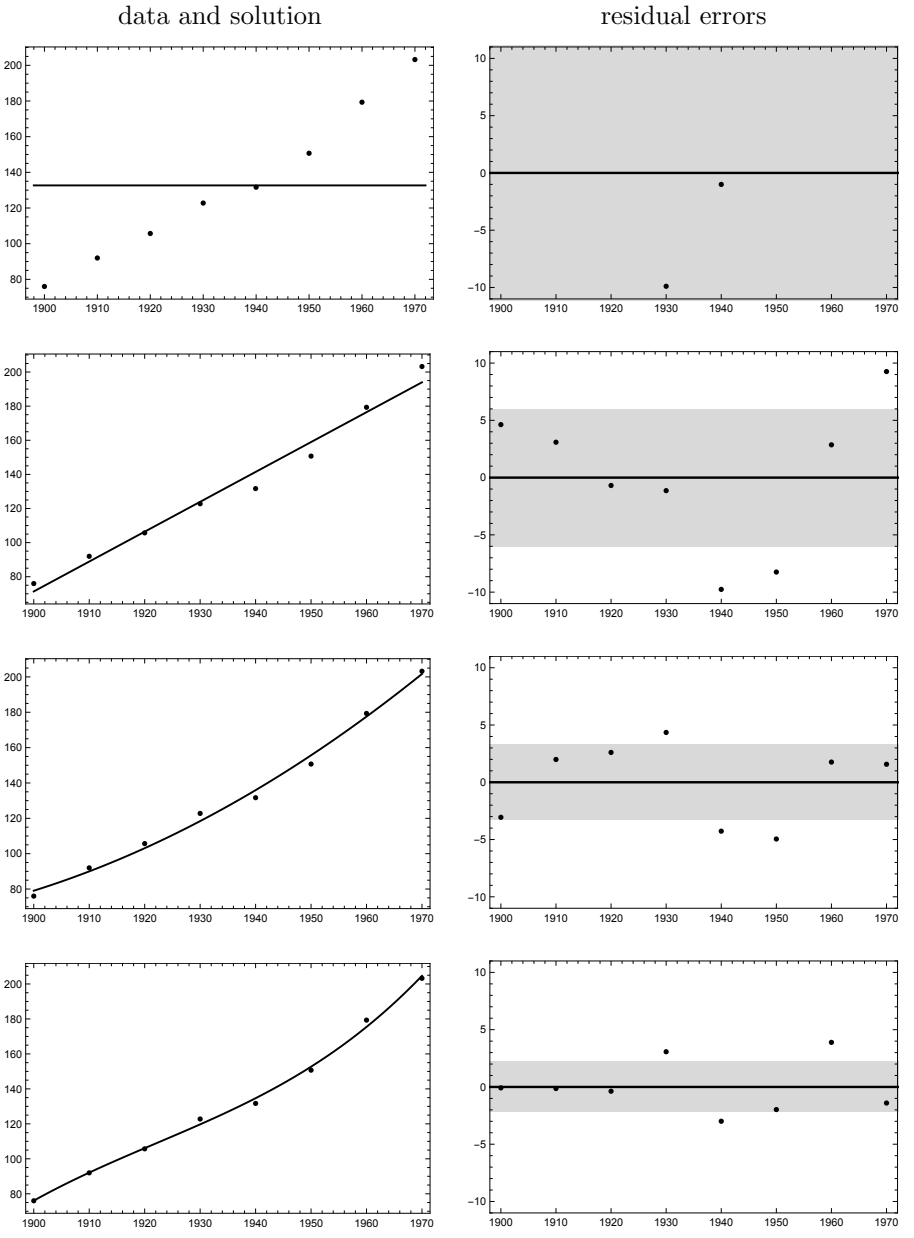
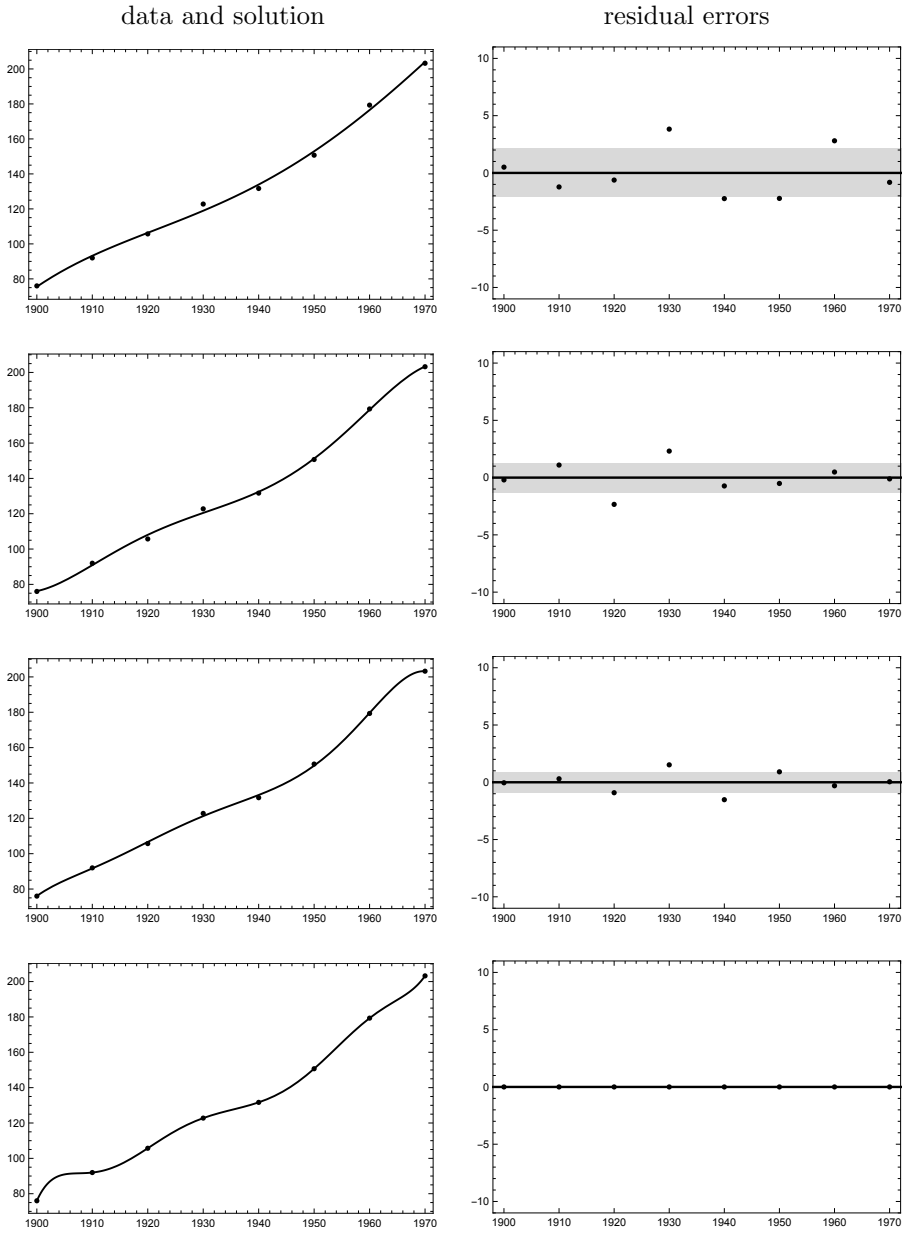


Table 15.7. *Fitting the census data with higher order polynomials.*



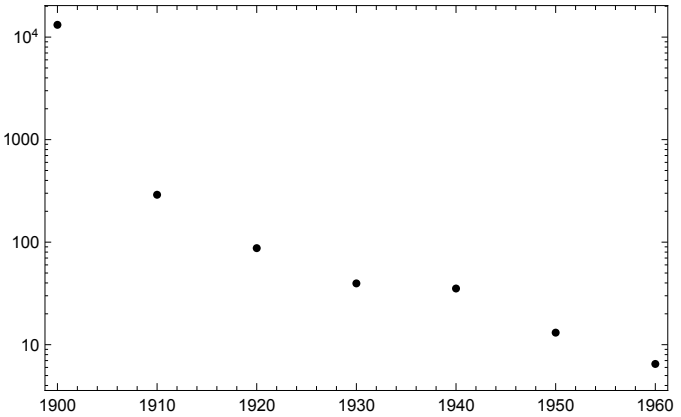


Figure 15.7. *Total error r*r by order of fit.*

Table 15.8. *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 V

Appendices

Appendix A

Exercises

A.1 Linear systems

§4.2.2 Equation (4.3) describes an extremum. How do we know this is a minimum, not a maximum? Because least squares solutions form a convex set. Proves this.

A.1.1 7.1.1

$$a_1 = \Delta^{-1}\left(\left(\mathbf{1}^{\mathrm{T}}\mathbf{1}\right)\left(x^{\mathrm{T}}T\right) - \left(\mathbf{1}^{\mathrm{T}}x\right)\left(\mathbf{1}^{\mathrm{T}}T\right)\right)$$

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

§7.1.1 Show that under the affine transformation $T \to \alpha T + \beta$, where $\alpha, \beta \in \mathbb{R}$, the slope transforms as

$$a_1 \to \alpha a_1.$$

By inspection, argue that a_0 is invariant.

Part VI

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