Chapter 1

Characteristics of Inverse Problems

The fields of inverse problems and statistics have significant overlap, though their literatures remain largely disjoint. Our goal in this manuscript is to present an introduction to the field of inverse problems with statistical content introduced, and connections between the two fields highlighted, wherever appropriate, in the hopes that this will facilitate the exchange of ideas between these two closely tied fields.

1.1 Preliminaries

We begin with the following linear statistical model with Gaussian noise:

$$\mathbf{b} = \mathbf{A}\mathbf{x} + \boldsymbol{\epsilon}, \tag{1.1}$$

where $\mathbf{b}$ is the $m \times 1$ model output, or observation, vector; $\mathbf{x}$ is the $n \times 1$ vector of unknown model parameters; $\mathbf{A}$ is the $m \times n$ forward model matrix, which sometimes requires a separate set of observations; and $\boldsymbol{\epsilon}$ is an $m \times 1$ independent and identically distributed (iid) Gaussian random vector with variance σ^2 across all entries, which we denote $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. We will assume that $m \geq n$ and that $\mathbf{A}$ has full column rank.

In the statistics literature, $\mathbf{b}$ is called the *response vector*, $\mathbf{A}$ the *design matrix*, and $\mathbf{x}$ the *unknown parameter vector*. Moreover, for statistical problems it is typically the case that $m \gg n$ and that the measured data consists both of the response vector $\mathbf{b}$ and of the elements of $\mathbf{A}$, which are called the *explanatory variables*.

The term *inverse problem* is used in many contexts, however, classically, inverse problems are *ill-posed*, which means not *well-posed*, a term first defined by Jacques Hadamard [10]. The definition, adjusted for our notation and situation, is given as follows.

Definition 1.1. *The discrete mathematical model* $\mathbf{b} = \mathbf{A}\mathbf{x}$ *is said to be* well-posed *if*

1. *for all model output* $\mathbf{b}$ *there exists a solution* $\mathbf{x}$ *to* $\mathbf{A}\mathbf{x} = \mathbf{b}$*;*

2. *the solution* $\mathbf{x}$ *is unique; and*

3. *the solution* $\mathbf{x}$ *is stable with respect to perturbations in* $\mathbf{b}$*; that is, if* $\tilde{\mathbf{b}}$ *is a perturbation of* $\mathbf{b}$ *with corresponding solution* $\tilde{\mathbf{x}}$*, then* $\|\mathbf{x} - \tilde{\mathbf{x}}\|/\|\mathbf{x}\|$ *is of the same order of magnitude as, or is smaller than,* $\|\mathbf{b} - \tilde{\mathbf{b}}\|/\|\mathbf{b}\|$*.*

It is difficult to make condition 3 precise in the discrete case and have it be applicable to all possible discrete inverse problems. This is not so in the continuous setting, where the focus is instead on the underlying continuously defined mathematical model $b = Ax$, of which $\mathbf{b} = \mathbf{A}\mathbf{x}$ is a numerical discretization. Here b and x are functions in a function space and A is an operator between function spaces; see [7, 27] for functional analytic treatments of inverse problems. In the continuous setting, Hadamard's condition 3 becomes

3. *the solution* x *is stable with respect to perturbations in* b*; that is, the inverse mapping* $b \mapsto x$ *is continuous.*

In an introductory course on linear algebra or statistics, students study linear regression, which often leads to a linear system $\mathbf{b} = \mathbf{A}\mathbf{x}$ that fails condition 1 of Definition 1.1. To see this consider the next example.

Example 1.2 Suppose as a scientist you want to find a formula to estimate the weight b in kilograms of a lion by using its length ℓ in meters. You have five samples: (b, ℓ) = (420,2.4), (350,2.0), (310,2.1), (280,1.8), and (75,1.3). We assume a model of the form

$$b_i = x_1 + x_2\ell_i + \epsilon_i, \tag{1.2}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Gathering the data in matrix form $\mathbf{b} = \mathbf{A}\mathbf{x}$ yields

$$\begin{bmatrix} 420 \\ 350 \\ 310 \\ 280 \\ 75 \end{bmatrix} = \begin{bmatrix} 1 & 2.4 \\ 1 & 2.0 \\ 1 & 2.1 \\ 1 & 1.8 \\ 1 & 1.3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \tag{1.3}$$

We plot this data in Figure 1.1. Visually, you can see that the data has a linear trend, but also that there is no single line that interpolates all data points. Or in other words, (1.3) does not have a solution. This fact can be verified by performing Gaussian elimination on the augmented system $[\mathbf{A}|\mathbf{b}]$. Thus (1.3) is an ill-posed problem because it fails conditions 1 and 2. We can overcome this deficiency by computing the least squares solution, or equivalently by solving the normal equations $\mathbf{A}^T\mathbf{A}\hat{\mathbf{x}} = \mathbf{A}^T\mathbf{b}$, yielding the best-fit line $b = -303.08 + 307.34\ell$, also plotted with the data in Figure 1.1. We will discuss least squares in more detail in a moment. ∎

The last example would not be considered a classical inverse problem. As was mentioned above, it is typical in inverse problems that the noise free equation $\mathbf{b} = \mathbf{A}\mathbf{x}$ in (1.1) is obtained from the numerical discretization of a continuously

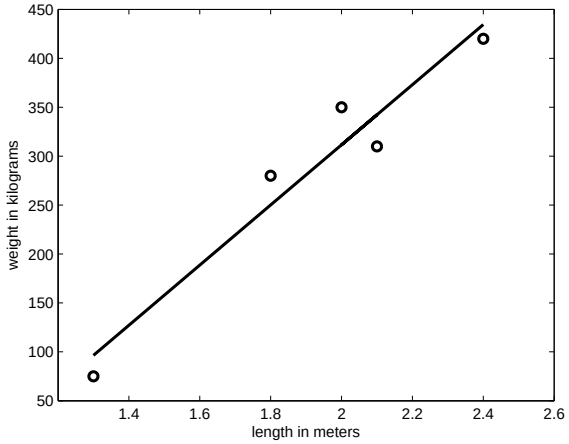

Figure 1.1. *Scatter plot of lion length and weight data* (ℓ_i, b_i) *together with a plot of the best fit line.*

defined physical model $b = Ax$, where b and x are functions and A is an operator between function spaces. In many important applications, and in the examples considered in this book, this continuous model takes the form of a Fredholm first kind integral equation

$$b(s) = Ax(s) \stackrel{\text{def}}{=} \int_\Omega a(s, s')x(s')\, ds', \tag{1.4}$$

where Ω is the domain over which x is defined, and a is known as the *kernel*. In the examples considered in this text, either $\Omega = [0, 1]$ or $\Omega = [0, 1] \times [0, 1]$.

To obtain a matrix-vector equation $\mathbf{b} = \mathbf{Ax}$ from (1.4), we must perform a numerical discretization. We assume $\Omega = [0, 1]$ and use the midpoint quadrature rule, which for a general function f defined on $[0, 1]$ has the form

$$\int_0^1 f(s')\, ds' = h \sum_{j=1}^{n} f(s'_j) + E_n, \tag{1.5}$$

where $h = 1/n$, $s'_j = (j - \frac{1}{2})/n$, and E_n is the quadrature error. Assuming the same grid for the s variable, i.e. $s_j = s'_j$, if we define $b_i \stackrel{\text{def}}{=} b(s_i)$, $x_j \stackrel{\text{def}}{=} x(s'_j)$, and $a_{ij} = a(s_i, s'_j)$, and apply (1.5) to (1.4), ignoring E_n, we obtain

$$b_i = h \sum_{j=1}^{n} a_{ij} x_j, \quad i = 1, \ldots, m. \tag{1.6}$$

Equation (1.6) can be written in matrix-vector form $\mathbf{b} = \mathbf{Ax}$, where $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{x} \in \mathbb{R}^n$, and $\mathbf{A} \in \mathbb{R}^{m \times n}$ with entries $[\mathbf{A}]_{ij} = a_{ij}$.

For many examples in imaging, $m = n$ and the matrix $\mathbf{A}$ is invertible. However, the resulting model $\mathbf{b} = \mathbf{A}\mathbf{x}$ will fail condition 3 of Definition 1.1, which will be the case for the next example.

Example 1.3 An application that we will repeatedly turn to in this manuscript, and that can be modeled as a Fredholm first kind integral equation, is image deblurring. In this section, we present a one-dimensional (1D) test problem. The ease of computation and visualization in the 1D case makes the illustration of certain concepts more straightforward. We will discuss the two-dimensional image deblurring problem later.

Often in image deblurring problems, the blur is the same everywhere in the image, in which case we say that it is spatially invariant. The assumption of a spatially invariant blur simplifies (1.4) so that it has convolution form

$$b(s) = \int_0^1 a(s - s')x(s')\, ds'. \tag{1.7}$$

Applying the mid-point quadrature rule to (1.7), as above, with kernel function

$$a(s) = \frac{1}{\sqrt{2\pi\gamma^2}} \exp\left(-\frac{s^2}{2\gamma^2}\right), \quad \gamma > 0, \tag{1.8}$$

yields the system of linear equations

$$b_i = h \sum_{j=1}^{n} a_{i-j} x_j, \quad i = 1, \ldots, m, \tag{1.9}$$

where $a_{i-j} = a((i-j)h)$. Compare with (1.6).

As with (1.6), equation (1.9) can be written as $\mathbf{b} = \mathbf{A}\mathbf{x}$ with $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{x} \in \mathbb{R}^n$, and $\mathbf{A} \in \mathbb{R}^{m \times n}$ with entries

$$[\mathbf{A}]_{ij} = \frac{h}{\sqrt{2\pi\gamma^2}} \exp\left(-\frac{((i-j)h)^2}{2\gamma^2}\right). \tag{1.10}$$

From this model, we can generate noisy data using (1.1) with $m = n = 80$ and σ^2 chosen so that the signal-to-noise ratio (SNR), $\|\mathbf{A}\mathbf{x}\|/\sqrt{n\sigma^2}$, is 50. Both $\mathbf{x}$ and $\mathbf{b}$ are plotted on the left in Figure 1.2.

∎

Example 1.4 The relationship between integration and it's inverse, differentiation, also illustrates a failure of condition 3 of Definition 1.1 nicely. The Second Fundamental Theorem of Calculus states that if

$$b(s) = \int_0^s x(s')\, ds', \tag{1.11}$$

then $\frac{d}{ds} b(s) = x(s')$. Unknown to most is that the inverse problem of obtaining x from b defined in (1.11) is ill-posed. To see this, let's proceed as in the last example

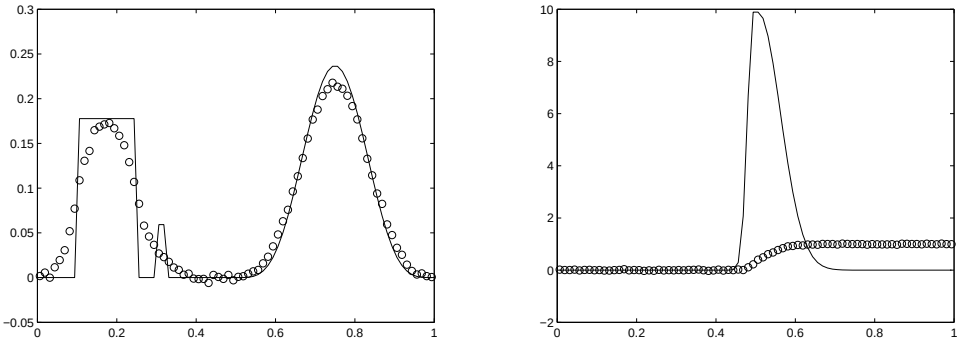

Figure 1.2. *One dimensional deblurring (left) and kernel reconstruction (right) examples. The true image* $\mathbf{x}$ *is given by the solid line and the blurred and noisy data* $\mathbf{b}$ *is given by the circles.*

and discretize (1.11) using midpoint quadrature. We have

$$b_i = \int_0^{s_i} x(s')\,ds' \approx h\sum_{j=1}^{i} x(s'_j), \quad i = 1, \ldots, n.$$

Letting $\mathbf{x}$ be the $n \times 1$ vector with jth entry $x(s'_j)$ yields $\mathbf{b} = \mathbf{A}\mathbf{x}$ with

$$\mathbf{A} = h\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \ddots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 1 & 1 & \cdots & 1 \end{bmatrix}_{n\times n}. \tag{1.12}$$

We will leave it as an exercise to show that

$$\mathbf{A}^{-1} = h^{-1}\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & \ddots & \ddots & 0 \\ 0 & -1 & 1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -1 & 1 \end{bmatrix}_{n\times n}. \tag{1.13}$$

Next, let's tie (1.11) to a specific application. Note that in the previous example, the kernel a was assumed to be known. However, in any image deblurring problem involving a kernel, an estimate of the kernel is needed before the inverse problem can be solved. In radiography, the kernel is estimated by passing a step function through the imaging instrument. If a convolution model is assumed, and

the step function has the form

$$\chi_t(s') = \begin{cases} 1 & s' > t, \\ 0 & s' \le t, \end{cases}$$

with $t \in [0, 1]$, then the output has the form (ignoring for a moment the finite computational domain):

$$\begin{aligned} b(s) &= \int_{-\infty}^{\infty} a(s - s')\chi_t(s')\, ds' \\ &= \int_{t}^{\infty} a(s - s')\, ds' \\ &= \int_{-\infty}^{s} a(s' - t)\, ds' \end{aligned}$$

Restricting to the computational domain [0,1], we use the model

$$b(s) = \int_0^s a(s' - t)\, ds',$$

which is equivalent to (1.11) with $x(s') = a(s' - t)$.

As a numerical test problem, we synthetically generates a kernel, which defines $\mathbf{x}$, and then, as in the previous example, generated synthetic data $\mathbf{b}$ with an SNR of 50 via (1.1) with $\mathbf{A}$ defined by (1.12). The result is plotted in Figure 1.2. ∎

1.2 The Least Squares Estimator

In the 1D examples above, the goal is to estimate the true image $\mathbf{x}$ given data $\mathbf{b}$ and the model matrix $\mathbf{A}$. A standard statistical approach is to compute the maximizer of the likelihood function defined by the linear model (1.1), which is given by

$$L(\mathbf{x} \,|\, \mathbf{b}) = \frac{1}{\sqrt{(2\pi)^n \sigma^{2n}}} \exp\left(-\frac{\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2}{2\sigma^2} \right). \tag{1.14}$$

The maximizer of $L(\mathbf{x} \,|\, \mathbf{b})$ is the minimizer of the negative log-likelihood $-\ln L(\mathbf{x} \,|\, \mathbf{b})$, or equivalently, of

$$\ell(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2, \tag{1.15}$$

and hence is also the least squares solution. The least squares solution can be expressed as the solution of the normal equations

$$\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}, \tag{1.16}$$

a fact whose proof we relegate to the exercises. We assume that $\mathbf{A}$ has full column rank, so that $\mathbf{A}^T\mathbf{A}$ is invertible and the least squares solution is given by

$$\mathbf{x}_{\text{LS}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}. \tag{1.17}$$

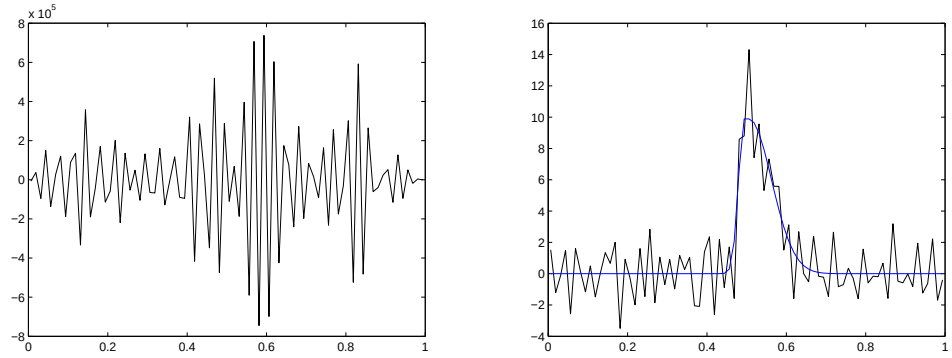

Figure 1.3. *The least squares solution* $\mathbf{x}_{\text{LS}}$ *for the one-dimensional deblurring example.*

Thus if $\mathbf{A}$ is invertible, then $\mathbf{x}_{\text{LS}} = \mathbf{A}^{-1}\mathbf{b}$.

The reader with any experience studying inverse problems will not be surprised that the least squares solutions for the above examples are poor. To illustrate, we plot the least squares solutions in Figure 1.3, and note that in both cases, highly oscillatory components degrade the reconstruction. In addition, for the deblurring example, the oscillatory components completely dominate, making the magnitude of the least squares solution more than 10^7 times larger than for the true image. To see that both examples fail condition 3 of Definition 1.1, in the deblurring case $\|\mathbf{Ax} - \mathbf{b}\|/\|\mathbf{Ax}\| = 0.024$ and $\|\mathbf{x} - \mathbf{x}_{\text{LS}}\|/\|\mathbf{x}\| = 5.25 \times 10^{13}$, while for the PSF reconstruction case $\|\mathbf{Ax} - \mathbf{b}\|/\|\mathbf{Ax}\| = 0.021$ and $\|\mathbf{x} - \mathbf{x}_{\text{LS}}\|/\|\mathbf{x}\| = 0.60$. Given these results, we say that the deblurring example is severely ill-posed, while the PSF reconstruction problem is only mildly ill-posed.

From a statistical viewpoint, we note that (1.1) implies $\mathbf{b} \sim \mathcal{N}(\mathbf{Ax}, \sigma^2\mathbf{I})$. Moreover, for a general $n \times 1$ normal random vector $\mathbf{v} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ if $\mathbf{w} = \mathbf{Bv}$, where $\mathbf{B}$ is an $m \times n$ matrix, $\mathbf{w} \sim \mathcal{N}(\mathbf{B}\boldsymbol{\mu}, \mathbf{BCB}^T)$. Thus from (1.17) we have

$$\mathbf{x}_{\text{LS}} \sim \mathcal{N}(\mathbf{x}, \sigma^2(\mathbf{A}^T\mathbf{A})^{-1}), \tag{1.18}$$

and hence, while the variance of each pixel of the data $\mathbf{b}$ is σ^2, for the i^{th} component of $\mathbf{x}_{\text{LS}}$ the variance is $[\sigma^2(\mathbf{A}^T\mathbf{A})^{-1}]_{ii}$, which will be large for large i in typical examples in inverse problems. As a result, the least squares solution will typically be far from the true image $\mathbf{x}$, which is again supported by the plots in Figure 1.3.

1.3 The Singular Value Decomposition and Ill-Conditioning

In order to understand why the least squares solutions are so poor, we turn to the singular value decomposition (SVD) of the design matrix:

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T, \tag{1.19}$$

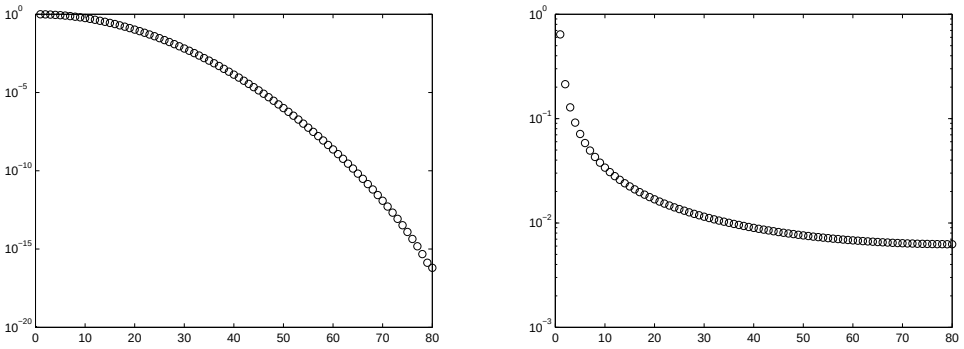

Figure 1.4. *A plot of singular values of* $\mathbf{A}$ *for the one-dimensional deblurring example (left) and the PSF reconstruction example (right). The y-axis is on the* log *scale.*

where $\mathbf{U}$ is an $m \times m$ matrix with columns given by the orthonormal eigenvectors of $\mathbf{A}\mathbf{A}^T$; $\mathbf{V}$ is an $n \times n$ matrix with columns given by the orthonormal eigenvectors of $\mathbf{A}^T\mathbf{A}$; and $\mathbf{\Sigma}$ is an $m \times n$ diagonal matrix with diagonal entries $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$, which are called the singular values of $\mathbf{A}$ and are the square roots of the eigenvalues of $\mathbf{A}^T\mathbf{A}$. The columns of $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_m]$ and $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_n]$ are called the left and right singular vectors of $\mathbf{A}$, respectively, and are orthonormal so that $\mathbf{U}^T\mathbf{U} = \mathbf{I}_m$and $\mathbf{V}^T\mathbf{V} = \mathbf{I}_n$.

Although $\mathbf{A}$ may not have an inverse, its pseudo-inverse always exists and is defined

$$\mathbf{A}^\dagger = \mathbf{V}\mathbf{\Sigma}^\dagger\mathbf{U}^T, \tag{1.20}$$

where $\mathbf{\Sigma}^\dagger$ is the $n \times m$ matrix with diagonal values $\sigma_1^{-1} \geq \sigma_2^{-1} \geq \cdots \geq \sigma_r^{-1} > 0$, where r is the smallest positive singular value. Detail on the SVD and the pseudo-inverse can be found in [9]. For discretized first kind Fredholm integral equations (called discrete inverse problems in [13]), the singular values decay continuously to 0 as i increases, with a large number clustered near 0. To illustrate, we plot the singular values of $\mathbf{A}$ in both cases in Figure 1.4. Note that the singular values drop to zero much more rapidly in the deblurring example, which partially explains the much worse results for the least squares solutions presented in Figure 1.3. The lower bound of the singular values in the simple integration (kernel reconstruction) example isn't much below the grid size h.

We can also say something about typical behavior of the left and right singular vectors, $\mathbf{u}_i$ and $\mathbf{v}_i$, respectively, for discretized first kind Fredholm integral equations; namely, they behave like the Fourier basis in the sense that as i increases, $\mathbf{u}_i$ and $\mathbf{v}_i$ become more oscillatory. This fact is demonstrated for several examples in [13]. We leave its verification for the 1D deblurring example considered above to the exercises.

1.3.1 Statistical Properties of the Least Squares Estimator

We now return to the least squares solution and view it in light of the above discussion. Given our assumptions regarding $\mathbf{A}$, it can be shown (see Exercise 8) that

$$\mathbf{x}_{\text{LS}} = \sum_{i=1}^{n} \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i} \mathbf{v}_i \tag{1.21}$$

$$= \mathbf{x} + \sum_{i=1}^{n} \frac{\mathbf{u}_i^T \boldsymbol{\epsilon}}{\sigma_i} \mathbf{v}_i. \tag{1.22}$$

From (1.22) and the above discussion, we see that due to the spectral properties of a typical noise realization $\boldsymbol{\epsilon}$, $(\mathbf{v}_i^T \boldsymbol{\epsilon})/\sigma_i$ will be large for large i, which serves to amplify the more oscillatory components in the second term in (1.22). This coincides with what we observe in the least squares solutions in Figure 1.3, where high frequency components dominate.

A more detailed analysis of the interplay between the noise $\boldsymbol{\epsilon}$, the data $\mathbf{b}$, and the spectral properties of $\mathbf{A}$, through its SVD, is possible using the discrete Picard condition; the interested reader should see [13] for details, as well as the exercises for a Picard analysis of the deblurring example considered in this chapter.

We can also use the SVD to shed further light on the distributional statement (1.18) for $\mathbf{x}_{\text{LS}}$. Replacing $\mathbf{A}$ with its the SVD in (1.18) yields

$$\mathbf{x}_{\text{LS}} \sim \mathcal{N}(\mathbf{x}, \sigma^2 \mathbf{V} \boldsymbol{\Sigma}^{-2} \mathbf{V}^T), \tag{1.23}$$

where $\boldsymbol{\Sigma}^{-2} = (\boldsymbol{\Sigma}^T \boldsymbol{\Sigma})^{-1}$.

From (1.23), we can see that $E[\mathbf{x}_{\text{LS}}] = \mathbf{x}$, and hence the least squares estimator is *unbiased*. However, in the orthonormal basis defined by the columns of $\mathbf{V}$ (the right singular vectors), (1.23) can be restated in a component-wise fashion as follows:

$$\mathbf{v}_i^T \mathbf{x}_{\text{LS}} \sim \mathcal{N}\left(\mathbf{v}_i^T \mathbf{x}, \sigma^2/\sigma_i^2\right), \quad i = 1, \ldots, n, \tag{1.24}$$

which directly shows that for $0 < \sigma_i \ll \sigma^2$ (i large), one can expect high variability in the corresponding oscillatory mode $\mathbf{v}_i$. Thus although the least squares estimator is unbiased, its *total variance*, which is given by

$$\text{tr}(\sigma^2 \mathbf{V} \boldsymbol{\Sigma}^{-2} \mathbf{V}^T) = \sigma^2 \sum_{i=1}^{n} 1/\sigma_i^2, \tag{1.25}$$

where 'tr' denotes the matrix trace, will be extremely large, which is undesirable.

Example 1.5 In order to make the above concepts more concrete, we consider a two dimensional example. First we define the matrix $\mathbf{A}$ via its SVD: let $\mathbf{v}_1 = [1/\sqrt{2}, 1/\sqrt{2}]^T$, $\mathbf{v}_2 = [-1/\sqrt{2}, 1/\sqrt{2}]^T$, and

$$\mathbf{A} = \mathbf{v}_1 \mathbf{v}_1^T + 10^{-2} \mathbf{v}_2 \mathbf{v}_2^T$$

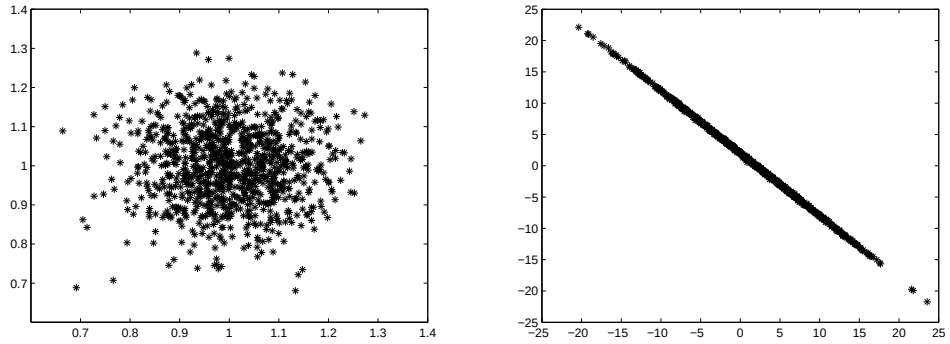

Figure 1.5. *On the left are 1000 realizations from the data model* $\mathbf{b} = \mathbf{A}\mathbf{x} + \boldsymbol{\epsilon}$*, and on the right are plots of the corresponding 1000 least squares solutions* $\mathbf{A}^{-1}\mathbf{b}$*. Both the true solution* $\mathbf{x}$ *and the noise-free data* $\mathbf{A}\mathbf{x}$ *are* $[1,1]^T$*.*

Note that we have used the summation form of the SVD (see Exercise 1.3). Hence $\mathbf{v}_1$ and $\mathbf{v}_2$ are both the left and right singular vectors and the singular values are $\sigma_1 = 1$ and $\sigma_2 = 10^{-2}$. Furthermore, as in the above cases, the smaller singular value corresponds to the 'higher frequency' vector. Finally, note that $\text{cond}(\mathbf{A}) = 100$, so that the matrix $\mathbf{A}$ is ill-conditioned.

Now suppose $\mathbf{x} = [1,1]^T$, then $\mathbf{b} = \mathbf{A}\mathbf{x} = [1,1]^T$ and $\mathbf{A}^{-1}\mathbf{b} = [1,1]^T$. However, if we add a small amount of noise to $\mathbf{b}$ using (1.1) with $\boldsymbol{\epsilon} = [0.026, 0.075]^T$ (a truncated realization of an iid Gaussian distribution with variance $\sigma^2 = 0.01$) then $\mathbf{A}^{-1}\mathbf{b} = [-1.400, 3.501]^T$, which is far from $[1,1]^T$.

Let us use the SVD discussion above to explore this problem more deeply. First, we plot 1000 realizations from the statistical model $\mathbf{b} = \mathbf{A}\mathbf{x} + \boldsymbol{\epsilon}$ with $\sigma^2 = 0.01$ on the left in Figure 1.5. Note that the $\mathbf{b}$ samples have a relatively small variance centered around $[1,1]^T$. Next, on the right in Figure 1.5, we plot $\mathbf{x}_{\text{LS}}$ for each realization of the data. Note that the noise free solution $[1,1]^T$ is contained within the data cloud, but that the variance is high in the direction of $\mathbf{v}_2$, while in the direction of $\mathbf{v}_1$ it is relatively small. This coincides with (1.24), which says that the variance of $\mathbf{x}_{\text{LS}}$ in the direction $\mathbf{v}_1$ is $\sigma^2/\sigma_1^2 = 0.01$, while in the direction $\mathbf{v}_2$ it is $\sigma^2/\sigma_2^2 = 100$. We computationally verified this for our experiment by computing the sample variance of $\mathbf{v}_1^T\mathbf{x}_{\text{LS}}$ and $\mathbf{v}_2^T\mathbf{x}_{\text{LS}}$ for all least squares solutions, which were 0.01 and 99.73, respectively. Additionally, note that

$$\mathbf{A} = \begin{bmatrix} 0.505 & 0.495 \\ 0.495 & 0.505 \end{bmatrix},$$

which, though non-singular, has columns that are 'nearly' linearly dependent, both well-approximating the singular vector $\mathbf{v}_1$. This agrees with our observations above since variance values for $\mathbf{v}_1^T\mathbf{x}_{\text{LS}}$ and $\mathbf{v}_2^T\mathbf{x}_{\text{LS}}$ of 0.01 and 100, respectively, suggest that the least squares solution is well-resolved in the direction of $\mathbf{v}_1$ and poorly resolved in the direction of $\mathbf{v}_2$. ∎

1.4 Summary

In this chapter, we presented a model that will appear throughout the remainder of this manuscript; namely the linear statistical model $\mathbf{b} = \mathbf{A}\mathbf{x}+\boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, where the noise-free equation $\mathbf{b} = \mathbf{A}\mathbf{x}$ arises from the numerical discretization of a first kind Fredholm integral equation (1.4).

To illustrate fundamental concepts, we introduced a one-dimensional image deblurring test case. We graphically demonstrated that for this problem the singular values of the design matrix $\mathbf{A}$ decrease continuously to zero, where they cluster, and that smaller singular values correspond to more oscillatory singular vectors (see Exercise 1.2.a). We then showed, using a SVD analysis, that when the least squares solution is computed, high frequency components of the noise in $\mathbf{b}$ are amplified, corrupting the solution. More specifically, assuming $(\sigma_i, \mathbf{v}_i)$ are the i^{th} singular value and right singular vector of $\mathbf{A}$, respectively, we showed that the variance of the least squares solution in the direction $\mathbf{v}_i$ is σ^2/σ_i^2 (see (1.24)).

We ended with an example containing only two unknowns that exhibited some of the same characteristics as linear systems arising from the discretization of Fredholm first kind integral equations, but which allowed for easy visualization and verification of the results of our SVD analysis.

Exercises

1.1. *This problem was written in the fall of 2011 by former University of Montana Math PhD students (and now PhD's) John Hossler, Marylesa Howard, and Jordan Purdy*: In baseball two commonly collected statistics for pitchers are batting average against (AVG) and walks plus hits per inning pitched (WHIP). Suppose we believe that a linear relationship exists between AVG and WHIP, and in particular, we believe we can predict a pitcher's WHIP from their AVG. AVG and WHIP data for 30 Major League Baseball (MLB) pitchers from the 2011 regular season (through September 2) are provided in `BaseballData.m`. Use this data to answer the following questions.

(a) Plot the batting average (AVG) against corresponding walks plus hits per inning pitched (WHIP) such that AVG predicts WHIP (AVG on x-axis, WHIP on y-axis).

(b) Write out by hand the first four rows of the design matrix. Construct the entire design matrix in MATLAB. Include an intercept in your model.

(c) Compute the least squares estimate. Note that $\boldsymbol{x}_{LS}$ is a 2×1 vector.

(d) Using the model created in part (c), plot the model on the data plot from part (a). This model predicts a pitcher's WHIP given his batting average against.

(e) For each observed data point, the residual is that observed value (WHIP) minus the corresponding predicted value (predicted WHIP) from the model. Calculate the thirty residuals and plot them against their corresponding AVG value. On the same graph, plot the line $y = 0$. Comment

on whether a pattern is distinguishable or whether there is random scatter in the residual plot. A pattern in the residuals may suggest a different model or a linear model with different error structure is more appropriate, for this data set. While random scatter does not verify our choice of this linear model is correct, it does suggest this model may be useful for analysis.

1.2. The midpoint quadrature rule for a function f defined on $[a, b]$ is defined by

$$\int_a^b f(x)dx = (b-a)f\left(\frac{a+b}{2}\right) + \frac{f''(\eta)}{24}(b-a)^3, \tag{1.26}$$

where $a \le \eta \le b$.

a. Write your own MATLAB code applying midpoint quadrature with $n = 100$ grid points to approximate $\int_0^{2\pi} \cos x\, dx = 1$.

b. Determine an upper bound for the quadrature error in part (a) using (1.26) (n times) and then verify that the inequality holds.

c. Modify `Deblur1d.m` so that it implements midpoint quadrature on the convolution model (1.7) with kernel

$$a(s) = \begin{cases} 100s + 10, & -\frac{1}{10} \le s \le 0, \\ -100s + 10, & 0 \le s \le \frac{1}{10}, \\ 0, & \text{otherwise.} \end{cases}$$

to obtain $\mathbf{b} = \mathbf{A}\mathbf{x}$. Plot a on $[-1, 1]$. Is the deblurring problem ill-posed with this kernel? Support your answer.

1.3. a. Show that $\mathbf{A}$ defined by (1.12) has inverse $\mathbf{A}^{-1}$ defined by (1.13). Use induction on n.

b. Show that $\mathbf{A}^{-1}$ corresponds to the discretization of $\frac{d}{ds}$ on the grid $\{s'_j\}_{j=1}^n$, where recall that $s'_j = (j - \frac{1}{2})/n$, with a zero Dirichlet boundary condition on the left (i.e., $s'_0 = 0$) and a zero Neumann boundary condition on the right (i.e., $s'_{n+1} = s'_n$). Use the forward difference approximation of the derivative:

$$\frac{d}{ds}x(s'_j) \approx \frac{x(s'_j) - x(s'_{j-1})}{h}.$$

c. Replace $\mathbf{A}$ by $\mathbf{A}^{-1}$ in `PSFrecon.m`. How do the results change? Is $\mathbf{b} = \mathbf{A}^{-1}\mathbf{x}$ an ill-posed problem, i.e. does it fail any of the three conditions of Definition 1.1?

1.4. A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if its Hessian $\nabla^2 f(\mathbf{x})$ is positive semi-definite for all $\mathbf{x}$ and is strictly convex if $\nabla^2 f(\mathbf{x})$ is positive definite for all $\mathbf{x}$. Moreover, the set of minimizers of a convex function f is the set on which the gradient of f is zero, i.e., those vectors $\mathbf{x}$ satisfying $\nabla f(\mathbf{x}) = \mathbf{0}$, and this set has a single member (and hence f a unique minimizer) if f is strictly convex. Use these facts to show that the minimizers of $\ell(\mathbf{x})$, with respect to $\mathbf{x}$, are the solutions of (1.16) and derive conditions on $\mathbf{A}$ that guarantee that $\ell(\mathbf{x})$ has a unique minimizer.

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

a. Compute the singular value decomposition of $\mathbf{A}$ by hand: $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$, where $\mathbf{U} \in \mathbb{R}^{3\times 3}$ is orthogonal, $\mathbf{S} \in \mathbb{R}^{3\times 2}$ is diagonal, and $\mathbf{V} \in \mathbb{R}^{2\times 2}$ is orthogonal.

b. Using the columns of $\mathbf{U}$ and $\mathbf{V}$ as basis elements, write down bases for the four subspaces: the column space $C(\mathbf{A})$, the row space $C(\mathbf{A}^T)$, the null space $N(\mathbf{A})$, and the left null space $N(\mathbf{A}^T)$.

c. Using the SVD you computed in (a), compute the pseudo-inverse $\mathbf{A}^\dagger$ of $\mathbf{A}$, and then use it to compute the least squares solution $\mathbf{x}_{\text{LS}} = \mathbf{A}^\dagger\mathbf{b}$ of

$$\begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

1. Compare the solution in (c) with the solution of the normal equations, $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$. The two solutions should agree.

1.6. For this problem, use the MATLAB m-files `Deblur1d.m` used to generate Figures 1.2-1.4.

a. Create plots verifying that the singular vectors of the design matrix $\mathbf{A}$ become more oscillatory as i increases.

b. A Picard plot for a discretized inverse problem is a plot of the values of σ_i, $|\mathbf{u}_i^T\mathbf{b}|$, and $|\mathbf{u}_i^T\mathbf{b}|/\sigma_i$. Picard plots are used extensively in [13] for analyzing discrete inverse problems. For the 1D deblurring example, create the Picard plot, first using the noise-free data and then using noisy data. Verify in both cases the *discrete Picard condition*: Ignoring the part of the Picard plot where $|\mathbf{u}_i^T\mathbf{b}|$ levels off due either to numerical round-off (in the noise-free case) or to the presence of noise in $\mathbf{b}$, the discrete Picard condition is satisfied if the remaining values of $|\mathbf{u}_i^T\mathbf{b}|$, on average, decay faster than σ_i.

c. Plot the variance values in (1.24). What do they tell you about the statistical properties of the least squares solution.

d. Show that even in the absence of noise, the least squares solution for the deblurring example remains poor. This is due to the fact that the deblurring example is so ill-conditioned that even round-off errors, which are on the order of machine precision (approximately 10^{-16}), are amplified.

1.7. Repeat Exercise 6 using `PSFrecon.m`.

1.8. a. Show that (1.19) can be equivalently written $\mathbf{A} = \sum_{i=1}^{n} \mathbf{u}_i\sigma_i\mathbf{v}_i^T$.

b. Substitute (1.19) into (1.17) and then show that (1.21) holds.

1.9. Let $\mathbf{v}$ be a random n-vector with mean $\boldsymbol{\mu} = E[\mathbf{v}]$ (here E is the expected value function) and covariance

$$\mathbf{C} = \mathrm{cov}(\mathbf{v}) = E[(\mathbf{v} - \boldsymbol{\mu})(\mathbf{v} - \boldsymbol{\mu})^T].$$

If $\mathbf{A}$ is an $m \times n$ matrix, the mean of $\mathbf{Av}$ is $E[\mathbf{Av}] = \mathbf{A}E[\mathbf{v}] = \mathbf{A}\boldsymbol{\mu}$. Use the linearity of the expected value function to show, as well, that $\mathrm{cov}(\mathbf{Av}) = \mathbf{ACA}^T$. Finally, use these results to prove (1.18).

1.10. There are two ways to sample from $\mathbf{x}_{\mathrm{LS}}$: (i) sample from (1.1) and compute $\mathbf{x}_{\mathrm{LS}}$ via (1.17); and (ii) sample directly from (1.18).

a. If $\mathbf{A}^{-1}$ exits, using the results from Problem 1.4 show that (1.18) is equivalent to $\mathbf{x}_{\mathrm{LS}} = \mathbf{x} + \sigma^2\mathbf{A}^{-1}\mathbf{v}$, where $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

b. In the m-file `TwoVarTest.m`, which was used to generate Figure 1.5, method (i) is used for sampling $\mathbf{x}_{\mathrm{LS}}$. Add a line of code implementing method (ii) and plot the two collection of samples together in different color to verify that the sample distributions look roughly the same. Also, compare their sample mean and covariance matrices using MATLAB's `mean` and `cov` functions.

Bibliography

[1] R. C. Aster, B. Borchers, and C. H. Thurber, *Parameter Estimation and Inverse Problems*, Elsevier, 2005.

[2] G. Backus and F. Gilbert, *The resolving power of gross earth data*, Geophysical Journal of the Royal Astronomical Society, **266** (1968), pp. 169-205.

[3] J. M. Bardsley, *An MCMC Method for Estimation and Uncertainty Quantification in Linear Inverse Problems*, submitted, Math Sciences, University of Montana, Tech. Report #28, 2010.

[4] J. M. Bardsley and J. A. Goldes, *An Iterative Method for Edge-Preserving MAP Estimation when Data-Noise is Poisson*, SIAM Journal on Scientific Computing, Vol.32, No.1, published online Feb 5, 2010.

[5] M. Bertero and P. Boccacci, *Inverse Problems in Imaging*, Institute of Physics, London, 1998.

[6] J. Besag, *Spatial Interaction and the Statistical Analysis of Lattice Systems*, Journal of the Royal Statistical Society, Series B, **36(2)** (1974), pp. 192-236.

[7] H. Engl, M. Hanke, A. Neubauer, *Regularization for Inverse Problems*, Springer 1996.

[8] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis, Second Edition*, Chapman & Hall/CRC, Texts in Statistical Science, 2004.

[9] G. Golub and C. Van Loan, *Matrix Computations, 3rd Edition*, Johns Hopkins University Press, 1996.

[10] J. Hadamard, *Sur les problmes aux drives partielles et leur signification physique*, Princeton University Bulletin, (1902), pp. 4952

[11] P. C. Hansen, *Analysis of discrete ill-posed problems by means of the L-curve*, SIAM Review, **34** (1992), pp. 561-580.

[12] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems*, SIAM, Philadelphia, 1996.

[13] P. C. Hansen, *Discrete Inverse Problems: Insight and Algorithms*, SIAM, Philadelphia, 2010.

[14] P. C. Hansen, J. G. Nagy, and D. P. O'Leary, *Deblurring Images: Matrices, Spectra, and Filtering*, SIAM, Philadelphia, 2006.

[15] Dave Higdon, *A primer on space-time modelling from a Bayesian perspective*, Los Alamos Nation Laboratory, Statistical Sciences Group, Technical Report, LA-UR-05-3097.

[16] J. Kaipio and E. Somersalo, *Statistical and Computational Methods for Inverse Problems*, Springer, 2005.

[17] D. Kincaid and W. Cheney, *Numerical Analysis: Mathematics of Scientific Computing*, Brooks/Cole, 2002.

[18] C. L. Mallows, *Some comments on C_P*, Technometrics, **15(4)** (1973), pp. 661-675.

[19] D. W. Marquardt, *Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation*, Technometrics, **12** (1970), pp. 591-612.

[20] R. H. Meyers, *Classical and Modern Regression with Applications*, 2nd ed., PWS Kent, Boston, 1990.

[21] J. Nagy and K. Palmer, *Steepst Descent, CG and Iterative Regularization of Ill-Posed Problems*, BIT, **43** (2003), pp. 1003-1017.

[22] J. Kamm and J. G. Nagy, *Kronecker Product and SVD Approximations in Image Restoration*, Linear Algebra and its Applications, **284** (1998), pg. 177–192.

[23] F. O'Sullivan, *A statistical perspective on ill-posed inverse problems*, Statistical Science, **1(4)** (1986), pp. 104-127.

[24] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications*, Chapman and Hall/CRC, 2005.

[25] L. Tenorio, *Statistical Regularization of Inverse Problems*, SIAM Review, **43** (2001), pp. 347-366.

[26] C. Van Loan, *The ubiquitous Kronecker product*, Journal of Computational and Applied Mathematics, **123** (2000), pp. 85-100.

[27] C. R. Vogel, *Computational Methods for Inverse Problems*, SIAM, Philadelphia, 2002.

[28] C. R. Vogel, *Non-convergence of the L-curve regularization parameter selection method*, Inverse Problems, **12** (1996), pp. 535-547.

[29] G. Wahba, *Practical approximate solutions to linear operator equations when the data are noisy*, SIAM Journal on Numerical Analysis, **14** (1977), pp. 651-667.