1 Introduction

The Kalman filter is probably the single most useful piece of mathematics developed in this century.
—John L. Casti, 2000 [Cas00]

1.1 Overview

This morning you awoke to a not unreasonable weather forecast. The last airplane you flew on may well have landed using a computer controlled autolander. These are examples of the power of modern control theory to both observe (weather) and control (airframes).

Such control techniques have several features. First, we embody our knowledge of a natural or manmade system into a mathematical computer model. It's hard to keep track in your head of atmospheric dynamics throughout the planet, just as it is hard to keep track of all the airframe control surfaces and dynamics of a large airplane. Our recent successes exhibit the confluence of several factors: We have increasingly powerful computers, we have increasingly powerful numerical methods to perform calculations, and we have better models. Boeing spent a fortune to develop the airframe model of the 777—it was an airframe developed almost entirely on computers. Similarly, we have a very good grasp of the physics of the atmosphere—the Euler equations [Kal03]; these convection equations are hard to solve, and one has to apply them in many places simultaneously to model the weather.

But building great models of systems is not enough.

First, *all of our models are wrong*. We are not interested in fitting data from the past; who cares that you could have predicted yesterday's weather given what you know today. Your model, born of past data, needs to interact with your system in real-time moving forward. You don't get a lot of data to work with—most sensor systems give a sparse sampling, in space and time, of the things you wish to measure. Many parts of the globe have little coverage of temperature, pressure, and wind speed, just as there are many portions of a modern airframe with no sensors nearby. So, if your model represents the whole system, it will need to reconstruct the parts that are inaccessible to measurement.

Second, *all measurements are bad*. Sensors are noisy, imprecise, and deteriorate, and the amplifiers change their properties with age, temperature, and calibration. So we need a way of optimally feeding our bad data to our wrong models.

Last, our *computers are never fast enough*. Nature runs along in continuous time, while our digital circuitry and computers chug along in saltatory bursts. Even if your model and sensor data were perfectly aligned at the last reading, the model's forward iteration takes place while nature does its own thing. A random air gust blows a plane sideways. Thermal stochastic effects, or chaotic sensitivity to conditions, create an atmospheric state whose trajectory differs from the one your model is creating. In a robotic system such as an airplane, the models are simple enough, and the computations fast enough, that many iterations can take place per second. In numerical weather forecasting, 6 hours is a typical window for North American and European agencies. And in both cases, we never use our best models. Speed is traded for accuracy—it does no good to take two days to predict tomorrow's weather.

There are two key concepts here: *observability* and *controllability*. Rudolf Kalman [Kal60] demonstrated that these concepts are linked: If you can observe the state of a system, you have just determined the extent to which you can control it. For all complex systems, theory and modeling is not a luxury or the province of the socially impaired. It is the only way you can observe such a system. It is the *lens* through which your observation changes from vague subjectivity to a comprehensive estimation of what you are observing.[1]

Airframes are, if my colleagues in aerospace engineering will forgive me, simple. We have also poured an incredible amount of funding into building very good models of these structures. The weather is complex, but compared with biological systems, it is simple. Gas and fluid dynamics are characterized by neighbor-to-neighbor interactions only, the molecules are all indistinguishable, and were it not for nonlinear dynamics such as turbulence, forecasting would be a breeze.

Brains are the most complex structures in the known universe. So on the one hand, you would never want to observe raw data without the benefit of modeling. But if all models are bad, models of brains are terrible.

This book is a gamble—a race between our gathering knowledge of neuroscience and its embodiment in computational neuroscience, and our growing sophistication in computational engineering tools that can handle the complexity of brain dynamics. At some point, the convergence of these two areas of knowledge will make the fusion of computational neuroscience and control theory absolutely required just to record from a neuron, or decipher an epileptic electroencephalogram (EEG). Are we there yet? Probably. The rest of this book will try to justify this statement.

[1] I was sure that this *lens* metaphor was suggested to me by Partha Mitra [MB08] in one of his talks, but neither he nor I are certain anymore. I will cite him more for inspiration if not fact.

1.2 A Motivational Example

The ultimate Figawi event.

In 1916, Ernest Shackleton was marooned with his crew in a place called Elephant Island after his ship sank in the Antarctic. They spent over a year there, and it looked like they were going to be trapped for another year as the winter again closed in. They had a photographer with them, and figure 1.1 shows the small lifeboat that they launched in an effort to traverse 800 miles across the worst ocean in the world to reach a tiny dot called South Georgia Island where there was a whaling station. By then the rest of the world had assumed that Shackleton, as had many other Antarctic explorers, been long dead. He took a sextant and a navigator, Frank Worsley, and a few strong men [Ale98].

Shackleton and Worsley had a near *perfect* model—a map of the position of land masses and oceans, and the navigational equations that could compute where Shackleton would be on a given day based on his previous position and his velocity. But perfect models, and deductive reckoning, can be terrible in implementation. Once he left Elephant Island his position grew increasingly uncertain. The only way to address this was to use additional data measurements to improve his estimation of position—his *state*. He needed to fuse new data, his position readings of the sun, with his existing model (map and estimates of position). He needed to *assimilate* data.

Shackleton had to take measurements at a time when the sky was clear and he could hold his sextant still. But in this part of the oceans, waves are often 40 feet high, and the wind

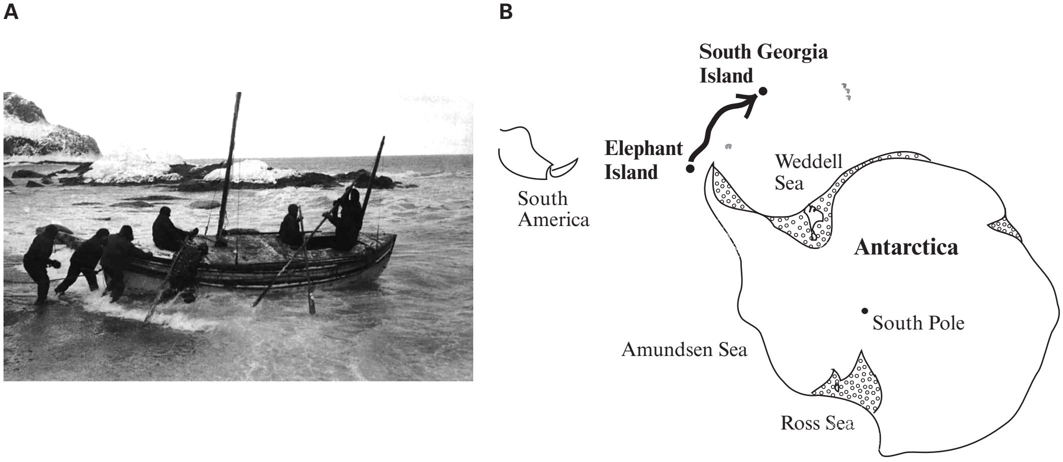

Figure 1.1
Voyage of the James Caird, 1916. Panel A reproduced from [Sha19], in public domain. Panel B courtesy of E. Schiff.

typically howls at 40 to 60 miles per hour. He was trying to grab a mast in the middle of the small boat to steady himself. The clouds broke only four times during the trip. There was enormous uncertainty about whether the position he measured in this turbulent world was accurate or not. Does he keep yesterday's prediction of where he expected to be that day, or use the measurement he just took? Compromising between model and data is a constant theme of this book. Flipping a coin to choose between estimate and data is not optimal—even bad measurements tell you *something*.

Data assimilation is a relatively new term forged out of two too common words. It is the fusion of data with your preexisting knowledge [WB07]. It is an act as old as navigation itself. Bayes's theorem (equation 1.73) tells you how to combine your previous guess with your new measurement, and this establishes a fundamental component of all data assimilation. But it would not be until the advent of the space program in the 1950s and 1960s that Kalman [Kal60] optimized how to apply Bayes's formula recursively to a dynamical process.

Let's simplify Shackleton's navigational nightmare a bit. Assume you are living in a one-dimensional world, and you only need to fix your position along a line (figure 1.2). Make the first measurement. I'm going to assume that when you made the first measurement, y, this is the first estimate of truth x. Not bad if you did this before pushing off Elephant Island—a good way to calibrate your instruments, or set your watch, because you really did know where you were and when the sun was at its noontime height. But perhaps your map of this rarely visited island did not fix its position well, or never captured its miserable coastline accurately. You might assume a probability distribution of your initial position x. Later in the book with nonlinear maps, you might pick a few initial conditions sprinkled

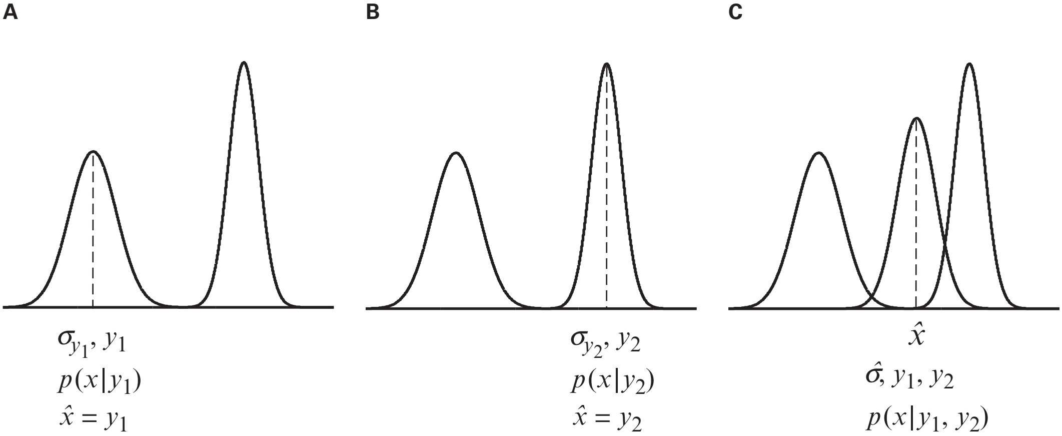

Figure 1.2
Bayes's rule tells you how to take sequential measurements and blend them. Two measurements, y_1 and y_2, and their respective uncertainties, σ_{y1} and σ_{y2}, are represented. The estimate of the true state, $\hat{x}$, is the mean of a probability distribution conditioned by y_1 or y_2, and then by y_1 and y_2.

about where you think you are, to determine what the consensus seems to be when you drop these initial guesses onto the rocky landscape of your topographical relief map.

Let's assume that you actually know something about the uncertainty, σ, for a given measurement, i, σ_i; that's quite a leap of faith,[2] but one that is often used in applications. Shackleton's first measurement, y_1, stinks, so as Caroline Alexander explained [Ale98], the rest of the crew helped to steady him and hold him up and reduce his weaving back and forth, for his second measurement, y_2. Assume that y_2 comes at a time very close to the first measurement. We don't think the boat has gone anywhere in the planet of significance during that short period of time, so we are estimating the same position on the map. Because two people were holding him up, the variance, the uncertainty of the second measurement, σ_{y_2}, is smaller than the first, σ_{y_1}. What is the optimal way to combine both measures to form an estimate, $\hat{x}$, of the true position, x?

There is a very critical, if subtle, truth here: *You never know the truth*. The truth is the true position x. You take measurements, y, and you use this information to *estimate* the truth, $\hat{x}$. It does not matter whether you are trying to estimate your position on a map, estimate a patient's pulse, asking what the voltage is within a neuron, or determining whether a seizure is going to occur soon. You have no direct access to the truth; you need to optimally estimate it given your knowledge (your model) and your data.

Shackleton's first measurement, y_1, had uncertainty σ_{y_1}, so if this were your only measurement, the conditional probability $p(x|y_1)$ of where you are, given what your measurement is (since you typically assume that your largest probability centers on your measurement), would lead you to pick $\hat{x}_1 = y_1$ as shown in figure 1.2A. Then you make the second measurement with its narrower uncertainty, as in figure 1.2B. If that were all you had, the conditional probability of x given only y_2 would be y_2. Now we combine these measurements into a single estimate (figure 1.2C).

There is no better intuitive explanation of this problem than that of Peter Maybeck, in his classic text [May82], and we will follow his framework here. Let's start with the answer (we will justify this later in the chapter):

$$\hat{x} = \frac{\sigma_{y_2}^2}{\sigma_{y_1}^2 + \sigma_{y_2}^2} y_1 + \frac{\sigma_{y_1}^2}{\sigma_{y_1}^2 + \sigma_{y_2}^2} y_2 \tag{1.1}$$

The estimated position, $\hat{x}$, is an average of the two position measurements, y_1 and y_2, weighted by the fraction of the variance from the *other* variances in equation (1.1). In other words, if the first measurement has a relatively large variance compared with the second, then pay attention to the second measurement more than the first (and vice versa). It also turns out that

$$\frac{1}{\hat{\sigma}_x^2} = \frac{1}{\sigma_{y_1}^2} + \frac{1}{\sigma_{y_2}^2} \tag{1.2}$$

[2] Kierkegarrd's input will not formally be required until the last chapter in this book.

which implies that

$$\hat{\sigma}_x^2 < \sigma_{y_1}^2 \text{ and } \hat{\sigma}_x^2 < \sigma_{y_2}^2 \tag{1.3}$$

Equation (1.3) develops a fundamental philosophical point: *all measurements are of value*. No matter how bad the variances $\sigma_{y_1}^2$ and $\sigma_{y_2}^2$ are, they *always* make $\hat{\sigma}_x^2$ smaller if you know them. If a measurement had infinite uncertainty, then equation (1.1) tells you to throw out its measurement, and equation (1.2) tells you to set the estimated uncertainty equal to the uncertainty of the better measurement. If there was ever no uncertainty in a measure, you just use it and stop taking more measurements. Perhaps this all seems intuitive, but equation (1.1) is a statement of weighted least squares, and it was Carl Gauss who originally found that these weights y's were optimal [Str86].

What if the uncertainties in these measures were all equal to each other? The estimate of position would be the average of the measurements. And the uncertainty $\hat{\sigma}_x^2$ would be *half* of the individual uncertainties. This is the case of ordinary (unweighted) least squares.

Ronald Fisher pointed out that there was an inverse relationship of uncertainty to the information content [Fis34, FC96]. If uncertainty is infinite, you have no information from that measurement. If uncertainty goes to zero, you know everything; were such a thing possible, you would not need Kalman filtering, or this book. Equation (1.2) tells you that information is always useful. You are never worse off by taking a measurement—not in this worldview, at least—as long as the measurement carries some information about what you're trying to estimate.

So far, we have made these measurements all at the same time, when more measurements always makes certainty better. We will shortly propagate that state through time, and then time will start to pull certainty apart.

Let's return to equation (1.1). Let's recognize that y_1 came before y_2, and say that y_1 is the information we have *prior* to y_2—our a priori knowledge. We will use the following trick, adding two terms that sum to zero, to eliminate the coefficient that multiplies the prior repeatedly in this book

$$\hat{x}_2 = \frac{\sigma_{y_2}^2}{\sigma_{y_1}^2 + \sigma_{y_2}^2} y_1 + \left\{ \frac{\sigma_{y_1}^2}{\sigma_{y_1}^2 + \sigma_{y_2}^2} y_1 \right\} + \frac{\sigma_{y_1}^2}{\sigma_{y_1}^2 + \sigma_{y_2}^2} y_2 - \left\{ \frac{\sigma_{y_1}^2}{\sigma_{y_1}^2 + \sigma_{y_2}^2} y_1 \right\} \tag{1.4a}$$

$$= y_1 + \frac{\sigma_{y_1}^2}{\sigma_{y_1}^2 + \sigma_{y_2}^2} (y_2 - y_1) \tag{1.4b}$$

$$= \hat{x}_1 + \frac{\sigma_{y_1}^2}{\sigma_{y_1}^2 + \sigma_{y_2}^2} (y_2 - \hat{x}_1) \tag{1.4c}$$

$$\hat{x}_2 = \hat{x}_1 + K(y_2 - \hat{x}_1) \tag{1.4d}$$

where

$$K = \frac{\sigma^2_{y_1}}{\sigma^2_{y_1} + \sigma^2_{y_2}} = \frac{\sigma^2_{\hat{x}_1}}{\sigma^2_{\hat{x}_1} + \sigma^2_{y_2}} \tag{1.5}$$

In (1.4b) the weighting is now applied to the difference $y_2 - y_1$. We assume in (1.4c) that when all you had was y_1, this was your best estimate of state $\hat{x}_1$. The weighting in (1.4d) is assigned the variable K, and in calculating K we can replace the a priori uncertainties σ^2_{y1} with $\sigma^2_{\hat{x}_1}$ as in equation (1.5).

K will shortly become Kalman's gain function [Kal60]. Kalman's gain, in this static case, is the ratio of the previous uncertainty to the total uncertainty. This ratio will tell you whether to weigh your model strongly using what you calculated a priori, or pay more attention to your new measurement, y_2. If $\sigma^2_{y_1}$ (or $\sigma^2_{\hat{x}_1}$) is small compared with the new uncertainty, ignore the new measure and "fly" the model. If your new measure is much more precise than the old one, you might want to forget your previous calculation and look out the window.

In *prediction-corrector* systems, you make a prediction, $\hat{x}_1$, and then correct it based on new measurements. The new prediction, $\hat{x}_2$, for linear systems with Gaussian errors, is the mean, the median, the mode, the maximum likelihood estimator, the weighted least squares error, the best linear unbiased estimator. All of this drops out of the simple formulation in equations (1.4).

K has been a guiding principle in water, air, and space navigation for over half a century. Let's take a look at how K relates to propagating the uncertainty.

From (1.2) we write

$$\frac{1}{\sigma^2_{\hat{x}_2}} = \frac{1}{\sigma^2_{y_1}} + \frac{1}{\sigma^2_{y_2}} = \frac{\sigma^2_{y_1} + \sigma^2_{y_2}}{\sigma^2_{y_1}\sigma^2_{y_2}} \tag{1.6}$$

so that

$$\sigma^2_{\hat{x}_2} = \frac{\sigma^2_{y_1}\sigma^2_{y_2}}{\sigma^2_{y_1} + \sigma^2_{y_2}} \tag{1.7}$$

We'll use our adding zero trick from equation (1.4a) again

$$\sigma^2_{\hat{x}_2} = \frac{\sigma^2_{y_1}\sigma^2_{y_2}}{\sigma^2_{y_1} + \sigma^2_{y_2}} - \left\{\frac{\sigma^2_{y_1}\sigma^2_{y_1}}{\sigma^2_{y_1} + \sigma^2_{y_2}}\right\} + \left\{\frac{\sigma^2_{y_1}\sigma^2_{y_1}}{\sigma^2_{y_1} + \sigma^2_{y_2}}\right\} \tag{1.8}$$

which resolves to

$$\sigma^2_{\hat{x}_2} = \sigma^2_{y_1} - K\sigma^2_{y_1} \tag{1.9}$$

or

$$\sigma^2_{\hat{x}_2} = \sigma^2_{\hat{x}_1} - K\sigma^2_{\hat{x}_1} \tag{1.10}$$

What happens to uncertainty when you make a second measurement? Uncertainty *always* goes down with a new measurement. You knew that from before, but in this framework, the Kalman gain tells you how to do a weighted average of not just position—it also lets you do a weighted adjustment of your ongoing knowledge of uncertainty. It's not obvious or trivial.

Now add dynamics. In the simplest case, the boat, after measurement y_2, moves. In Shackleton's case, the wind probably blows with a very nice constant mean velocity, v, say 50 miles per hour all day and all night long, but there are some random fluctuations, q, which we will add to our constant velocity

$$\frac{dx}{dt} = velocity + noise = v + q, \qquad q = N(0, \sigma_q^2) \tag{1.11}$$

where $N(0, \sigma_q^2)$ indicates that q is drawn from a normal (Gaussian) distribution, with mean $=$ 0 and variance $= \sigma_q^2$.

Integrate this equation to make your next position prediction. We integrate the position, velocity, and uncertainty from time $t = t_2$ to $t = t_3$

$$\int_{t_2}^{t_3} dx = \int_{t_2}^{t_3} vdt + \int_{t_2}^{t_3} qdt \tag{1.12}$$

which gives

$$x_3 - x_2 = v(t_3 - t_2) + \sigma_q^2(t_3 - t_2) \tag{1.13}$$

The integration of position and velocity gives the trivial results $x_3 - x_2$ and $v(t_3 - t_2)$. The integration of the random q is extremely nontrivial: $\sigma_q^2(t_3 - t_2)$. Intuitively, you can sense why this is true. If you integrate a Brownian noise process, that is integrate values drawn from a Gaussian distribution throughout a time interval, the result happens to be the variance times that time interval. This is known in physics as a Langevin equation (a particle diffusing with the wind also blowing), and in economics it is related to the Black-Scholes equation (a way to price options). Klebaner's textbook [Kle05] is an excellent resource to seek further understanding of such stochastic integrals.

In the previous case, from t_1 to t_2, there was no mean velocity, and if we had waited a given amount of time, the boat's position would diffuse on the map from random gusts. The uncertainty in position increases uniformly with time for the case of static estimation.

We will need to introduce some new notation to clarify two very different time scales that will be referred to throughout this book. There is a slow time scale that represents the natural system's dynamics over an interval, say t_2 to t_3, or the analogous time computed from the predictive model such as equation (1.12). In contrast, there is the much faster time scale that represents the state of a system just before you make a new measurement, $\hat{x}_3^-$, and the advance in your knowledge just after you take in a new measurement, $\hat{x}_3^+$. Indeed, even the computation required to absorb this new piece of data, as when you use equation (1.4d), is faster than model propagation as with the integration in equation (1.12). And in our boat example, the boat has not gone any appreciable distance just as you make a measurement—but you have learned something afterwards.

So we take our best estimated position at time t_2, $\hat{x}_2$, and the estimated uncertainty, $\hat{\sigma}_q^2$, and propagate both forward with equation (1.13). You now have a best predicted position, $\hat{x}_3^-$. There is the additive integrative uncertainty, $\sigma_q^2(t_3 - t_2)$, which if we permitted too long a time interval, would generate planetary scale uncertainty. So we don't want to take measurements too far apart.

Now take y_3, the next measure. Shackleton could complete only four measurements during the journey [Ale98]. What is the new position and uncertainty?

If the uncertainty of the new measurement is too high, the Kalman gain goes to zero, and by equation (1.4d), you ignore y_3. If the uncertainty is very small, then just use y_3 and ignore the model. Kalman filtering gives you a prescription for what data to use and what data to ignore based on the uncertainties of either of your measurements or the underlying process. But we have considerable ground yet to cover before discussing the Kalman filter in detail.

If we focus on the mean from the example in equation (1.13) (the mean of σ_q^2 is zero), the mean propagates as

$$\hat{x}_3^- = \hat{x}_2^+ + v\,(t_3 - t_2) \tag{1.14}$$

and the variance as

$$\sigma^2_{\hat{x}_3^-} = \sigma^2_{\hat{x}_2^+} + \sigma_q^2\,(t_3 - t_2) \tag{1.15}$$

all before the measurement at $t = t_3$ is taken.

Now make measurement y_3 with variance $\sigma^2_{y_3}$. To assimilate this new measurement, we write

$$\hat{x}_3^+ = \hat{x}_3^- + K_3(y_3 - \hat{x}_3^-) \tag{1.16}$$

where we now index K acknowledging our new time scale as

$$K_3 = \frac{\sigma^2_{\hat{x}_3^-}}{\sigma^2_{\hat{x}_3^-} + \sigma^2_{y_3}} \tag{1.17}$$

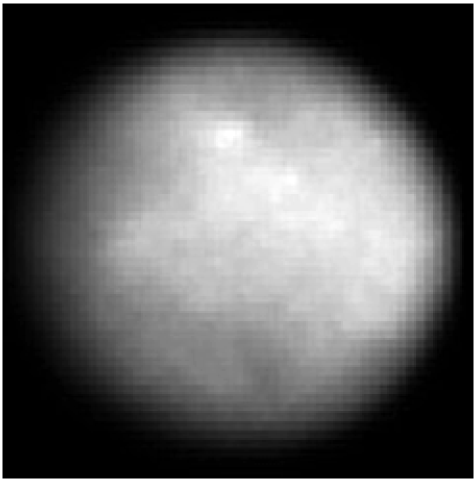

Figure 1.3
The minor planet Ceres—the largest asteroid. Image from Hubble Space Telescope.

and

$$\sigma^2_{\hat{x}_3^+} = \sigma^2_{\hat{x}_3^-} - K_3\sigma^2_{\hat{x}_3^-} \tag{1.18}$$

Note carefully the subtleties of how increasing or decreasing $\sigma^2_{y_3}$ or $\sigma^2_{\hat{x}_3^-}$ affects K_3, and the implications of $\sigma^2_{y_3} \ll \sigma^2_{\hat{x}_3^-}$, or $\sigma^2_{\hat{x}_3^-} \ll \sigma^2_{y_3}$.

1.3 Least Squares

But since all our measurements and observations are nothing more than approximations to the truth, the same must be true of all calculations resting on them, and the highest aim of all computations made concerning concrete phenomena must be to approximate, as nearly as practicable, to the truth.
—Carl Frederic Gauss, 1809

At the end of the eighteenth century, astronomers had sighted the minor planet Ceres (now relegated to largest asteroid status, figure 1.3). In his 1809 book [Gau09] on tracking asteroids and comets,[3] Gauss detailed that he had been using his least squares method since 1795 (when he was 18), much to the distress of his contemporary Legendre [Sor70]. Gauss worked out how to solve least squares and weighted least squares solutions to deal with the measurement of celestial objects. His was the only technique that seemed capable of

[3] Henry Davis's very readable translation [GD57] of Gauss's 1809 book, *Theory of the Motion of Heavenly Bodies*, is now readily available through the Google book project on the Internet, and well worth examining.

completing the elliptical orbit and estimating where Ceres was going to be after it transited behind the sun for a period of weeks, using only the incomplete measurements of previous sightings before the transit [TW99].

Gauss, by 1809, had laid down most of the foundations for later measurement theory, Kalman filtering, and this book. His insights included conceptualizing that one needed at least an approximate knowledge of the dynamics of a system in order to assimilate data, and that errors in measurements required measuring more than the minimum number of observations needed in order to solve the equations in your possession (i.e., your problem should be *overdetermined*) [Sor70]. He realized that one needed to minimize the errors between what you predicted and what you observed—what will later be called the *innovations* in modern control theory. And he invented and understood the far-reaching properties of least squares:

> [T]he most probable system of values of the unknown quantities . . . in which the sum of the squares of the differences between the observed and computed values of the functions is a minimum. . . . This principle, which promises to be of most frequent use in all applications of the mathematics to natural philosophy, must, everywhere, be considered an axiom with the same propriety as the arithmetical mean of several overserved values of the same quantity is adopted as the most probable value. [GD57]

Let's introduce the matrix formalism of Gauss's least squares. Throughout this book, we wish to solve the basic problem

$$Ax = y \tag{1.19}$$

where y is our observation and, if y is a vector, then A is a matrix.[4] We need matrices because we assume that all of our natural (and neural) systems have more than one variable that describes their states x, and that there will be more than one variable measured in any observation y. Even in this world of multielectrode arrays, one still might record from, for instance, a single deep brain microelectrode, but you would always consider that univariate y to be a condensation of a multivariable (multiple neuron) state. The matrix A tells us how to *remix* x, that is form a linear combination of the actual underlying variables that make up the vector or matrix x, in order to form the measurement y. We always assume that the true state x of any system, neural or otherwise, is hidden from our direct observation. Our task is to find the best (most probable) estimate of x given observations y, and we call this best estimate $\hat{x}$.

As a simple example, let's assume that y is measured twice, with values 1 and 3, and that x is one-dimensional with a coefficient $a = 1$ at each measurement time, then[5]

[4] We will use capital letters to indicate matrices in the text.

[5] If you are unfamiliar with matrix mathematics and linear algebra, I would strongly recommend putting this book down and spending a few weeks with one of Gilbert Strang's introductory linear algebra textbooks [Str06]. These are well suited for self-instruction for the neuroscientist or physician who did not realize that such skills would be needed later in life. Note, also, that MIT has made the video recordings of Professor Strang's course on this subject openly available on the MIT Open Courseware Web site.

$$\begin{bmatrix}1\\1\end{bmatrix} x = \begin{bmatrix}1\\3\end{bmatrix} \tag{1.20}$$

You do this every day in a laboratory or clinic—make up for imprecision by repeating measurements. Such systems are termed *overdetermined*. But if the measurements differ, and the system is assumed to remain the same, *none* of the individual measurements is correct. Some overdetermined systems are famous examples of great science; Millikan reported measuring the electron charge on his oil drop 58 times on 60 consecutive days—the result was fundamental and permanent [McG71].[6] We define the vector of errors, r, as

$$\text{error} = r = Ax - y = \begin{bmatrix}x-1\\x-3\end{bmatrix} \tag{1.21}$$

The length of a vector is indicated by the *norm* $\|\cdot\|$, and following the suggestion of Gauss, we wish to minimize the square of the norm of r

$$\|r\|^2 = r_1^2 + r_2^2 + \ldots = [Ax-y]^T[Ax-y] \tag{1.22}$$

where $[\cdot]^T$ indicates matrix transpose of rows and columns. The square of the norm is found by taking the inner product of the transpose of the error vector with itself. To minimize it, take the derivative and set it equal to zero

$$\frac{d}{dx}\left[(x-1)^2 + (x-3)^2\right] = 2(x-1) + 2(x-3) = 0 \tag{1.23}$$

which tells you that the best estimate of x is $\hat{x} = 2$. In matrix notation, this would be

$$\frac{d}{dx}\left[x^T A^T Ax - y^T Ax - x^T A^T y + y^T y\right] = 0 \tag{1.24}$$

Taking derivatives of each term with respect to x yields[7]

$$2A^T Ax - 2A^T y = 0 \tag{1.25}$$

so we reduce to one equation with one unknown

[6] Some considerable controversy has arisen in recent years over these measurement reports. See [Goo00], for instance.

[7] The derivative of the quadratic form $x^T A^T Ax$ is a bit confusing at first sight. For vectors that are functions of x, $u(x)$ and $v(x)$, the derivative with respect to x of $u^T v$ is

$$\frac{d}{dx}\left[u^T v\right] = \left[\frac{du}{dx}\right]^T v + \left[\frac{dv}{dx}\right]^T u$$

So write $Ax = u = v$, and express $x^T A^T Ax$ as $(Ax)^T(Ax) \equiv u^T v$, and the derivative is seen to readily be $2A^T Ax$.

$$A^T A x = A^T y \tag{1.26}$$

Using our example from equation (1.20), we get

$$[1\ 1] \begin{bmatrix} 1 \\ 1 \end{bmatrix} x = [1\ 1] \begin{bmatrix} 1 \\ 3 \end{bmatrix} \tag{1.27}$$

where $2x = 4$, and we find our best estimate, $\hat{x} = 2$, as we already knew. So in matrix formalism, the solution to least squares for such problems is

$$\hat{x} = (A^T A)^{-1} A^T y \tag{1.28}$$

Our discussion above was for the case where each measurement had equal uncertainty. Now let's introduce *weighted* least squares, where we assume, as in our Shackleton story, that the measurements have different uncertainties.

Strang [Str86], provides an excellent discussion of weighted least squares, and how they lead to the Kalman filter, and we will follow that discussion closely here. The weights will be indicated by matrix W, so that

$$W A x = W y \tag{1.29}$$

and using our example from equation (1.20)

$$\begin{bmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} x = \begin{bmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix} \tag{1.30}$$

The σ's are standard deviations. Our weighted error vector is now

$$\text{error} = W r = W A x - W y \tag{1.31}$$

and we need to minimize

$$\|Wr\|^2 = \sigma_{11}^2 r_1^2 + \sigma_{22}^2 r_2^2 + \ldots \tag{1.32}$$

which leads to

$$A^T W^T W A x = A^T W^T W y \tag{1.33}$$

The best estimator, $\hat{x}$, is

$$\hat{x} = (A^T W^T W A)^{-1} A^T W^T W y \tag{1.34}$$

Equation (1.34) is unpleasant, but using our previous example, we write

$$[1\ 1] \begin{bmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{bmatrix} \begin{bmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} x = [1\ 1] \begin{bmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{bmatrix} \begin{bmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

which yields

$$(\sigma_{11}^2 + \sigma_{22}^2)x = (\sigma_{11}^2 \cdot 1 + \sigma_{22}^2 \cdot 3)$$

so our best estimate, $\hat{x}$, is

$$\hat{x} = (\sigma_{11}^2 \cdot 1 + \sigma_{22}^2 \cdot 3)/(\sigma_{11}^2 + \sigma_{22}^2) \tag{1.35}$$

If the weightings are all equal, we can set all the σ_{ii}'s to 1, and $W = I$, the *identity* matrix with all 1's on the diagonal and zeros elsewhere. This would bring us back to ordinary least squares. Otherwise,

$$\hat{x} = \frac{\sigma_{11}^2}{(\sigma_{11}^2 + \sigma_{22}^2)} y_1 + \frac{\sigma_{22}^2}{(\sigma_{11}^2 + \sigma_{22}^2)} y_2 \tag{1.36}$$

which explains the form of our first lifeboat-inspired example[8] in equation (1.1).

1.4 Expectation and Covariance

We need to develop a few more essential concepts. The *central limit theorem* states[9] that the sum of many random processes gives a Gaussian probability distribution $p(x)$,

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-x^2/2\sigma^2\right] \tag{1.37}$$

where we define that the total probability must equal 1

$$\int_{-\infty}^{\infty} p(x)dx = 1 \tag{1.38}$$

The *expectation* of x, the mean μ_x of what you would expect after a large (infinite) number of samples, is defined as

$$E[x] = \sum_x xp(x) = \mu_x \tag{1.39}$$

for discrete, and

[8] Note that the indices used in W, σ_{ij}, are used to label the rows i and columns j in the matrix, and are not the indices of the weightings from equation (1.1), where σ_{11} was σ_{y_2}.

[9] The proof of the central limit theorem is not at all trivial [KF70], but most readers will be content if it is just stated as done here.

$$E[x] = \int_{-\infty}^{\infty} xp(x)dx \tag{1.40}$$

for continuous processes. For our Gaussian probability distribution with zero mean

$$E[x] = 0 \tag{1.41}$$

Variance is the expectation of the squared deviation from the mean (for discrete)

$$\mathrm{var}[x] = E[(x - E[x])^2] = \sum_x (x - \mu_x)^2 p(x) \equiv \sigma_x^2 \tag{1.42}$$

The variance for the continuous Gaussian distribution is

$$E[x^2] = \int_{-\infty}^{\infty} x^2 p(x)dx \tag{1.43}$$

The *covariance*[10] is the expectation of the squared deviation from the mean for multiple variables

$$\mathrm{cov}[x, y] = E[(x - \mu_x)(y - \mu_y)] = \sum_x \sum_y (x - \mu_x)(y - \mu_y)p(x, y) \tag{1.44}$$

where $p(x, y)$ is the joint probability distribution of x and y. If x and y are vectors, then covariance is the expectation of the outer product

$$\mathrm{cov}[x, y] = E[(x - \mu_x)(y - \mu_y)^T] \tag{1.45}$$

So for an error vector

$$\mathrm{r} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \end{bmatrix} \tag{1.46}$$

the covariance matrix is

$$E[rr^T] = R \tag{1.47}$$

Gauss showed [Str86] that the inverse of the measurement covariance matrix, R^{-1}, gives the *best linear unbiased estimator* for the least squares solution, so we replace $W^T W$ in equation (1.34) with R^{-1}

[10] A superb introduction to covariance and multivariable statistics in general can be found in Bernhard Flury's textbook [Flu97].

$$\hat{x} = (A^T R^{-1} A)^{-1} A^T R^{-1} y \tag{1.48}$$

There are now two sets of errors to keep track of for the rest of the book. R is the measurement error covariance matrix, the measurement errors being $Ax - y$. The other set of errors that concern us are the errors in the estimation of x, which are $x - \hat{x}$. We define P as the covariance in the errors in the estimation of x

$$P = E[(x - \hat{x})(x - \hat{x})^T] \tag{1.49}$$

Once we know the result in equation (1.48), it follows that the best estimate of P is

$$P = (A^T R^{-1} A)^{-1} \tag{1.50}$$

The proof is nontrivial, and the least unpleasant description of it is in Strang (p. 144 in [Str86]).

If one refers back to figure 1.2c, R describes the errors in measurements y, and P describes the errors in the estimate of $\hat{x}$.

Let's assume that you take several measurements of a neuron firing rate. Assume for now that the uncertainty in measurements, σ^2, are all the same and independent from each other. Then the covariance matrix is just the set of individual variances

$$R = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \tag{1.51}$$

and the inverse of a diagonal matrix is just the reciprocal of the diagonal values

$$R^{-1} = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/\sigma^2 \end{pmatrix} \tag{1.52}$$

Then from $P = (A^T R^{-1} A)^{-1}$,

$$P^{-1} = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/\sigma^2 \end{pmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{1.53}$$

which is just

$$P = \frac{\sigma^2}{2} \tag{1.54}$$

and for n measurements

$$P = \frac{\sigma^2}{n} \tag{1.55}$$

P decreases with every measurement. It does not matter how uncertain the measurements are—all measurements tell you something about the truth x.

1.5 Recursive Least Squares

Solving for $\hat{x}$ or P using equation (1.48) or (1.50) gets more complex as the number of measurements and the size of matrices A and R enlarge. But each new measurement changes the covariances R and P only incrementally. If the errors

$$\begin{bmatrix} r_1 \\ r_2 \\ \vdots \end{bmatrix} \tag{1.56}$$

are independent, then for two measurements at t_0 and t_1 (again, following [Str86])

$$\mathrm{R} = \begin{bmatrix} R_0 & \\ & R_1 \end{bmatrix} \tag{1.57}$$

and

$$P^{-1} = \begin{bmatrix} A_0^T & A_1^T \end{bmatrix} \begin{bmatrix} R_0^{-1} & \\ & R_1^{-1} \end{bmatrix} \begin{bmatrix} A_0 \\ A_1 \end{bmatrix} \tag{1.58}$$

Multiply this out and you get

$$P^{-1} = A_0^T R_0^{-1} A_0 + A_1^T R_1^{-1} A_1 \tag{1.59}$$

which is the key to making a recursive formula. Using equation (1.50) we can write

$$P_1^{-1} = P_0^{-1} + A_1^T R_1^{-1} A_1 \tag{1.60}$$

and this holds for any sequential estimations.

Substituting equation (1.50) in (1.48)

$$\hat{x}_1 = P_1 A^T R^{-1} y \tag{1.61}$$

which, for our simple two-measurement example, yields

$$\hat{x}_1 = P_1 \begin{bmatrix} A_0^T & A_1^T \end{bmatrix} \begin{bmatrix} R_0^{-1} & \\ & R_1^{-1} \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \end{bmatrix} \tag{1.62}$$

which multiplied out gives

$$\hat{x}_1 = P_1 \left[P_0^{-1} A_0^{-1} y_0 + A_1^T R_1^{-1} y_1 \right] \tag{1.63}$$

Substituting x_0 for $A_0^{-1} y_0$

$$\hat{x}_1 = P_1 \left[P_0^{-1} x_0 + A_1^T R_1^{-1} y_1 \right] \tag{1.64}$$

and finally substituting P_0^{-1} with equation (1.60)

$$\hat{x}_1 = P_1 \left[P_1^{-1} x_0 - A_1^T R_1^{-1} A_1 x_0 + A_1^T R_1^{-1} y_1 \right] \tag{1.65}$$

gives

$$\hat{x}_1 = x_0 + K_1 (y_1 - A_1 x_0) \tag{1.66}$$

with

$$K_1 = P_1 A_1^T R_1^{-1} \tag{1.67}$$

Equations (1.66) and (1.67) are the beginning steps of *recursive least squares*. Note several things. If $y_1 = A_1 x_0$, then $\hat{x}_1 = x_0$. If $y_1 = A_1 x_0 + e$, where e is the unexpected part of y_1, the *innovation*, then $\hat{x}_1 = x_0 + K_1(e)$, where $K_1(e)$ is the correction to the previous x_0.

The fundamental equations of recursive least squares are therefore

$$\begin{aligned} P_i^{-1} &= P_{i-1}^{-1} + A_i^T R_i^{-1} A_i \\ K_i &= P_i A_i^T R_i^{-1} \\ \hat{x}_i &= \hat{x}_{i-1} + K_i (y_i - A_i \hat{x}_{i-1}) \end{aligned} \tag{1.68}$$

Let's measure the firing rate of a neuron in spikes per minute by counting spikes within a 10-second window.[11] You get one, two, and four spikes, with calculated $y_0 = 6$, $y_1 = 12$, and $y_2 = 24$ spikes per minute. We assume $A = 1$. Then

$$\begin{aligned} A^T A \hat{x} &= A^T y \\ \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \hat{x} &= \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 6 \\ 12 \\ 24 \end{bmatrix} \\ 3\hat{x} &= 42 \\ \hat{x} &= 14 \text{ spikes per minute} \end{aligned} \tag{1.69}$$

and

[11] Strang [Str86] shows a similar example for estimating heart rate.

$$P^{-1} = \begin{bmatrix} A_0^T & A_1^T & A_2^T \end{bmatrix} \begin{bmatrix} R^{-1} & & \\ & R^{-1} & \\ & & R^{-1} \end{bmatrix} \begin{bmatrix} A_0 \\ A_1 \\ A_2 \end{bmatrix}$$

$$P^{-1} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1/\sigma^2 & & \\ & 1/\sigma^2 & \\ & & 1/\sigma^2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \tag{1.70}$$

$$P = \frac{\sigma^2}{3}$$

The best estimate of this static problem (we assume that this pacemaker neuron always has the same rate) is 14 spikes per minute.

Now let's estimate this firing rate recursively:

$$P_0^{-1} = A_0^T R^{-1} A_0 = [1][1/\sigma^2][1] = 1/\sigma^2$$

$$P_1^{-1} = P_0^{-1} + A_1^T R^{-1} A_1 = 1/\sigma^2 + 1/\sigma^2 = 2/\sigma^2 \tag{1.71}$$

$$P_2^{-1} = P_1^{-1} + A_2^T R^{-1} A_2 = 3/\sigma^2$$

with

$$K_n = P_n A_n^T R_n^{-1} = \frac{\sigma^2}{n}[1]\frac{1}{\sigma^2} = \frac{1}{n}$$

$$\hat{x}_1 = \hat{x}_0 + K_n(y_1 - \hat{x}_0) = 6 + \frac{1}{2}(12 - 6) = 9 \tag{1.72}$$

$$\hat{x}_2 = 9 + \frac{1}{3}(24 - 9) = 14$$

Note that in the recursive formulation that you throw away all of the history with each step—there is no need to store steadily increasing amounts of data with such recursion. Three measurements are not so bad to calculate. But later, we will have data sets with many thousands of sequential samples, and if you could find a computer with enough memory to solve equation (1.69), it would not be able to keep up with tracking a process in real time. Equations (1.71) and (1.72) never get more complex than what you see.

1.6 It's a Bayesian World

If our data were y, and our true underlying states x, then one could describe probability distributions of y independent of x, $p(y)$, and distributions of x independent of y, $p(x)$.

But if x and y are related, from knowledge of $p(x)$, we could refine the distribution of y given $p(x)$, the *conditional distribution* $p(y|x)$. But we already have y—this book is about estimating the underlying truth x, and often about estimating the most likely x, the mean of x, from the data observed. The conditional distribution $p(x|y)$ is the *posterior* or a posteriori distribution of x given data y, and $p(x|y)$ is our primary concern.

In taking conditional expectations, one takes a slice of a *joint* probability distribution, $p(x, y)$, and since all probability distributions must add to one as in (1.38), we need to normalize things. The rule for this comes from *Bayes's rule* [Flu97]

$$p(y|x) = \frac{p(x, y)}{p(x)} \tag{1.73}$$

So instead of integrating over $p(x)$ as in (1.40), the conditional expectation $E[x|y]$ is

$$E[x|y] = \sum_x x p(x|y) = \sum_x x \frac{p(x, y)}{p(y)} = \mu_{x|y} \tag{1.74}$$

and with similar structure for the continuous version. Multiply two (independent) Gaussian probability distributions as in equation (1.37) to get the joint Gaussian distribution

$$p(x, y) = \frac{1}{\sqrt{2\pi\sigma_x^2\sigma_y^2}} \exp\left[-x^2/2\sigma_x^2 - y^2/2\sigma_y^2\right] \tag{1.75}$$

and the conditional distribution $p(y|x)$ is constructed as [Flu97]

$$p(y|x) = \frac{1}{\sqrt{2\pi\sigma_{y|x}^2}} \exp\left[-(y - \mu_{y|x})^2/2\sigma_{y|x}^2\right] \tag{1.76}$$

So let's assume that the truth is normally distributed with a mean of μ_x and variance σ_x^2, and that the observations y are normally distributed with a mean of $\mu_{y|x}$ and variance $\sigma_{y|x}^2$. Then, following [WB07],

$$p(y|x) = \frac{1}{\sqrt{2\pi\sigma_{y|x}^2}} \exp\left[-(y_1 - \mu_{y|x})^2/2\sigma_{y|x}^2\right] \cdot \frac{1}{\sqrt{2\pi\sigma_{y|x}^2}} \exp\left[-(y_2 - \mu_{y|x})^2/2\sigma_{y|x}^2\right] \tag{1.77}$$

which is proportional to

$$\exp - \left[(y_1 - \mu_{y|x})^2/2\sigma_{y|x}^2 + (y_2 - \mu_{y|x})^2/2\sigma_{y|x}^2\right] \tag{1.78}$$

To simplify this a bit, we can assume that the real position x is known, and the y's are distributed about the true position x, which serves as the y mean. Bayes's rule tells you that $p(x|y)$ is proportional to $p(y|x)p(x)$, so we multiply (1.78) by $p(x)$ to get

$$\exp - \left[(y_1 - x)^2/2\sigma_{y|x}^2 + (y_2 - x)^2/2\sigma_{y|x}^2 + (x - \mu_x)^2/2\sigma_x^2\right] \tag{1.79}$$

which, because the sum of y^2 terms can be factored out (they are the variance of y plus the mean of y, all constants), simplifies to

$$\exp - \left[x^2\left(\frac{2}{\sigma_{y|x}^2} + \frac{1}{\sigma_x^2}\right) - 2x\left(\frac{y_1 + y_2}{\sigma_{y|x}^2} + \frac{\mu_x}{\sigma_x^2}\right)\right] \tag{1.80}$$

and after completing the square,[12] this new Gaussian distribution has a mean of

$$E[x|y] \equiv \hat{x} = \frac{\sigma_x^2}{2\sigma_x^2 + \sigma_{y|x}^2}(y_1 + y_2) + \frac{\sigma_{y|x}^2}{2\sigma_x^2 + \sigma_{y|x}^2}\mu_x \tag{1.83}$$

or for n measurements, letting $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$,

$$E[x|y] \equiv \hat{x} = \frac{n\sigma_x^2}{n\sigma_x^2 + \sigma_{y|x}^2}\bar{y} + \frac{\sigma_{y|x}^2}{n\sigma_x^2 + \sigma_{y|x}^2}\mu_x \tag{1.84}$$

Notice that these fractions, weighting the means of the distributions by the fraction of the *other* distribution's variance, is the reason why such weighting was used in equation (1.1) at the start of this chapter. It was based on Bayes's rule.

Note some important subtleties of equation (1.84). If the uncertainty in the model, σ_x^2, grows large, you can ignore the model—it's a bad map. If the number of measurements, y_i, becomes large, the data overwhelms your prior information, μ_x and σ_x^2, and you can again throw out the model. If $\sigma_x^2 \to 0$, ignore the measurements. Last, note that if we rewrite equation (1.84) as

$$\hat{x} = \mu_x + \frac{n\sigma_x^2}{n\sigma_x^2 + \sigma_{y|x}^2}(\bar{y} - \mu_x) \tag{1.85}$$

[12] If $ax^2 + bx = 0$, then

$$x^2 + \frac{b}{a}x + \left(\frac{b}{2a}\right)^2 = \left(\frac{b}{2a}\right)^2 = \left(x + \frac{b}{2a}\right)^2 \tag{1.81}$$

where

$$\frac{b}{2a} = -\frac{\sigma_x^2\sigma_{y|x}^2}{n\sigma_x^2 + \sigma_{y|x}^2}\left[\frac{y_1 + y_2}{\sigma_{y|x}^2} + \frac{\mu_x}{\sigma_x^2}\right] \tag{1.82}$$

for n measurements.

and let

$$K = \frac{n\sigma_x^2}{n\sigma_x^2 + \sigma_{y|x}^2} \tag{1.86}$$

we have the more general form of equation (1.4d).[13]

Kalman filtering is a subset of Bayesian analysis. But Kalman added dynamics to this static data assimilation framework. And that's the subject of the next chapter.

Exercises

1.1. The inverse of a matrix can be calculated in several ways [Str06]. One approach is to use the determinant of a matrix and calculate a set of matrix cofactors, where

$$(A^{-1})_{ij} = \frac{C_{ji}}{\det(A)}$$

and the cofactors, C_{ij}, are

$$C_{ji} = (-1)^{i+j} M_{ij}$$

where, for a 2 × 2 matrix,

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

the M's are matrix minors,[14] which for this 2 × 2 case are

$$M_{11} = a_{22},\ M_{22} = a_{11},\ M_{12} = a_{12},\ M_{21} = a_{21}$$

and the determinant is $a_{11}a_{22} - a_{12}a_{21}$. The inverse is

$$\frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$$

Suppose you are recording from two neurons, and measure their firing rates as $y_1 = 20$, and $y_2 = 30$. Given a linear system of equations

$$Ax = y$$

[13] Unresolved in this chapter, or history in general, is just how Shackleton and Worsley actually found South Georgia Island. Armed with the techniques discussed in this book, you have a better chance. Armed with the technology of the turn of the twentieth century, their feat stands as a remarkable human achievement.

[14] A matrix minor is the determinant after deleting the row and column containing a_{ij}.

where

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

solve by hand the best least squares solution for

$$\hat{x} = (A^T A)^{-1} A^T y \tag{1.87}$$

1.2. Using the same A and y from Exercise (1.1), solve the least squares solution to (1.87) using a computer.[15] Note that, although you can solve this in Matlab-Octave as `xhat = inv(A'*A)*A'*y`, that a set of more numerically stable algorithms is accessed using the notation `xhat = (A'*A)\A'*y`. If the inverse of `(A'*A)` does not exist, the pseudoinverse can be tried as `xhat = pinv(A'*A)*A'*y`.

1.3. Another computer exercise. Assume the measured firing rates of a neuron are

$$y = \begin{bmatrix} 3 \\ 5 \\ 4 \\ 8 \end{bmatrix}$$

spikes per second. If, as in equation (1.69), we assume that

$$A = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

solve for $\hat{x}$ using

$$\hat{x} = (A^T A)^{-1} A^T y$$

Assuming that

$$R = \begin{bmatrix} \sigma^2 & & & \\ & \sigma^2 & & \\ & & \sigma^2 & \\ & & & \sigma^2 \end{bmatrix}$$

[15] Throughout this text, I will sketch out algorithms compatible with Matlab and Octave. Although Matlab is a commonly used language for scientific computing, it is expensive. Octave is an open source equivalent, and I will strive to ensure that any algorithmic examples are code compatible between these languages.

calculate P from

$$P = (A^T R^{-1} A)^{-1}$$

Now repeat the calculation using the fundamental equations (1.68) of recursive least squares

$$\begin{aligned}
P_i^{-1} &= P_{i-1}^{-1} + A_i^T R_i^{-1} A_i \\
K_i &= P_i A_i^T R_i^{-1} \\
\hat{x}_i &= \hat{x}_{i-1} + K_i (y_i - A_i \hat{x}_{i-1})
\end{aligned}$$