Estimation I

Ian Reid
Hilary Term, 2001

1 Introduction

Estimation is the process of extracting information about the value of a parameter, given some data related to the parameter. In general the data are assumed to be some random sample from a "population", and the parameter is a global characteristic of the population.

In an engineering context, we are often interested in interpreting the output of a sensor or multiple sensors: real sensors give inexact measurements for a variety of reasons:

- Electrical noise – robot strain gauge;
- Sampling error – milling machine encoder (see figure 1);
- Calibration error – thermocouple
- Quantization/Shot noise – CCD (see figure 2)

We seek suitable mathematical tools with which to model and manipulate uncertainties and errors. This is provided by **probability theory** – the "calculus of uncertainty".

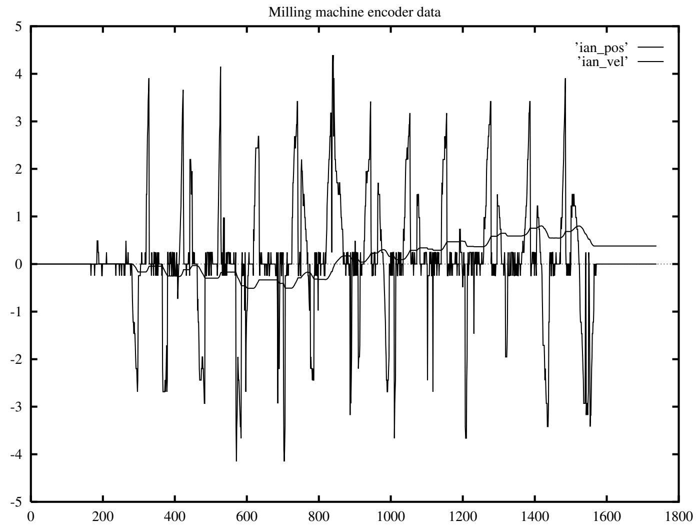


Figure 1: Position and velocity information from a milling machine encoder.

The following (by no means exhaustive) list of texts should be useful:

- Bar-Shalom and Fortmann, "Tracking and Data Association", Academic Press, 1988.
- Brown, "Introduction to Random Signal Analysis and Kalman Filtering", Wiley, 1983.

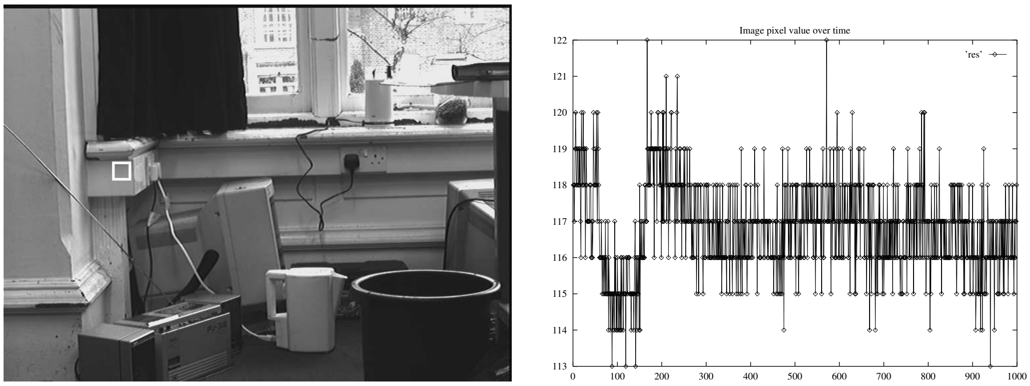


Figure 2: Pixel data over time for a stationary camera.

- Gelb (ed), “Applied Optimal Estimation”, MIT Press, 1974.
- Jacobs, “Introduction to Control Theory”, 2nd Edition, OUP, 1993.
- Papoulis, “Probability and Statistics”, Prentice-Hall.
- Press, Flannery, Teukolsky, and Vetterling, “Numerical Recipes in C”, CUP, 1988.
- Riley, Hobson and Bence, “Mathematical Methods for Physics and Engineering”, CUP, 1998.

1.1 Modelling sensor uncertainty

Consider a sensor which measures a scalar quantity, and taking one measurement. We could use this measurement as an **estimate** of the value of the sensor – just about the simplest estimate we could think of.

However, in general the measurement will not exactly equal the true value. Rather it will be displaced from the true value because of the effects of noise, etc.

An appropriate way to model this is via a **probability distribution**, which tells us how likely particular measurement is, given the true value of the parameter. We write this as $P(Z|X)$, the probability that Z is observed, given that X is the true state being measured.

In a general case we may have that Z is not a direct observation, and X may be a function of time. In a simple case, we may know measurement must lie within ϵ of the true value, but no more – i.e the sensor model can be described by a **uniform** distribution (see figure 3)

$$p_Z(z|x) \sim U(x - \epsilon, x + \epsilon)$$

1.2 Modelling uncertain prior information

In addition to modelling a sensor as a probability distribution in order to capture the notion of uncertainty in the measurements, we can apply the same idea to modelling uncertain prior information, since the parameter(s) to be estimated may have known statistical properties.

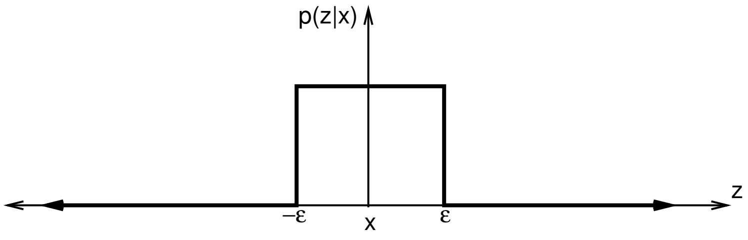


Figure 3: Modelling a sensor by $p(\text{Observation}|\text{True state})$ uniformly distributed around the true value.

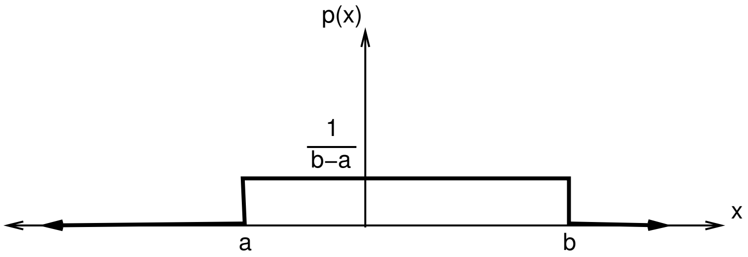


Figure 4: A uniform prior

A simple example, once again, is where we have no knowledge of the parameter, other than that it lies within a certain range of values. These could be fixed by known physical constraints – for example, the pressure of a gas in a cylinder is known to lie between $0kPa$ and some upper limit set by the strength of the cylinder, or the temperature of some molten iron in a furnace lies between 1535^oC and 2900^oC. We could also use a uniform distribution here

$$p_X(x) \sim U(a, b)$$

(see figure 4)

Of course we may know more about the parameter; the temperature of water in a kettle which is switched off will lie somewhere between the ambient room temperature and 100^oC, depending on how recently it was used, but is more *likely* to be close to room temperature in the overall scheme of things.

1.3 Combining information

We require some way of combining a set of measurements on a parameter (or parameters) and/or some prior information into an estimate of the parameter(s).

Two different ways of combining will be considered:

Summation (figure 5(a)) Consider timing multiple stages of a race (e.g. RAC Rally). Given some uncertainty model for each of the stage timings, how can we determine an uncertainty model for the total time – i.e. the sum of a set of random variables?

Pooling (figure 5(b)) Now consider using two different watches to time the same stage of the race.

Can these two measurements be combined to obtain a more accurate estimate of the time for the stage.

(a)

(b)

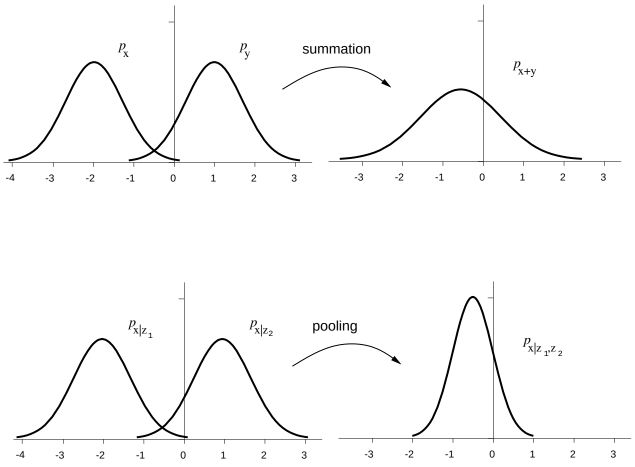


Figure 5: Two ways of combining information.

In the latter case, an average of the two might seem like an appropriate choice. The **sample mean** of a set of observations $z_i, i = 1 \ldots n$, is given by

$$\hat{x} = \frac{1}{n} \sum_{i=1}^{n} z_i$$

This is a **linear estimator**, since it is a linear combination of the observations. It can be shown (see later) that it has a smaller variance than any of the individual measurements, so in that sense it is potentially a better estimator of the true value x.

Note that this estimator has not taken into account any prior information about the distribution for x. If $p(x)$ is available, we can use **Bayes' rule**

$$p(x|z) = \frac{p(z|x)p(x)}{p(z)}$$

to combine measurements $p(z|x)$ and the **prior** $p(x)$, as in figure 6, obtaining the **posterior**.

1.4 Example: Navigation

Consider making an uncertain measurement on the location of a beacon, followed by an uncertain motion, followed by another noisy measurement. Combine these to obtain an improved estimates of the beacon's and your own location – see figure 7.

1.5 Example: Differential GPS

GPS is the Global Positioning System. With a simple receiver costing a few hundred pounds you can obtain an estimate of your latitude/longitude to within $\pm 100m$ from a network of satellites which

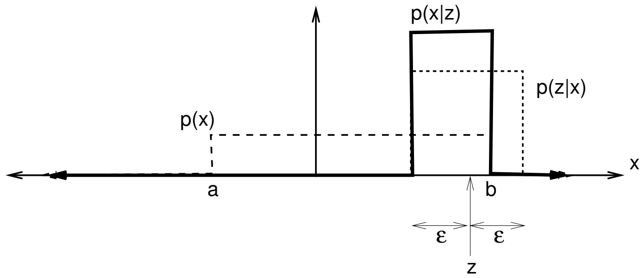
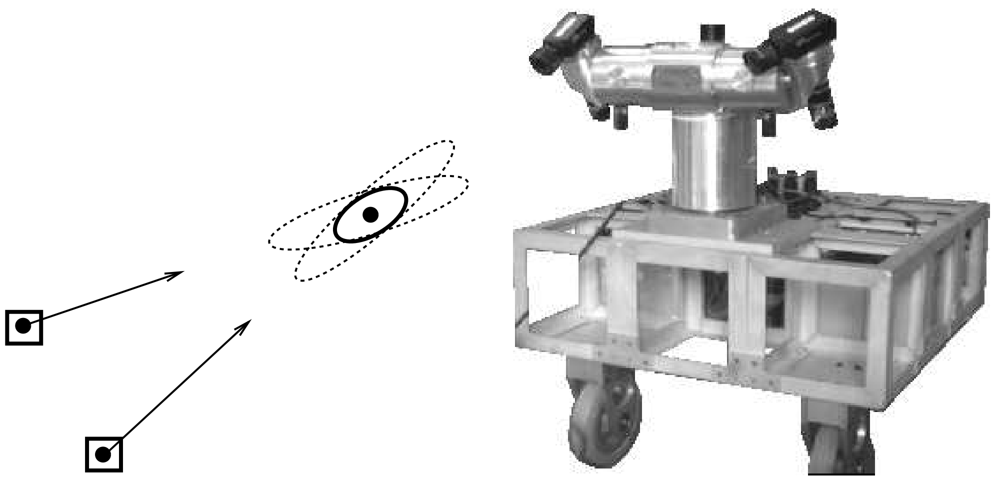


Figure 6: Combining the prior $p(x)$ with a measurement $p(z|x)$ to obtain the posterior



(a) (b)

Figure 7: Multiple views of a beacon to improve its location estimate (a) conceptually; (b) application to autonomous mobile robot navigation.

cover the globe. The system has accuracy down to a few metres for the military, but for civilian use the signal is deliberately corrupted by noise. The noise is:

- identical for all locations for a given instant in time;
- highly correlated for closely spaced time intervals (i.e. slowly changing);
- effectively uncorrelated for time instants widely separated in time.

[Exercise: think about what the auto-correlation function and power spectral density might look like...]

Hence by keeping a unit in one place for a long period of time and averaging the position data, you can form a much more accurate estimate of where you are.

The idea of DGPS is to have two receivers which communicate. One is mobile, and the other is kept stationary, hence its position is known rather more accurately than the mobile one. However at any one instant, the "noise" on the two estimates is perfectly correlated – i.e. identical – so the position estimate for the mobile unit can be significantly improved from the $\pm 100m$ civilian spec to that of the stationary unit. See figure 8.

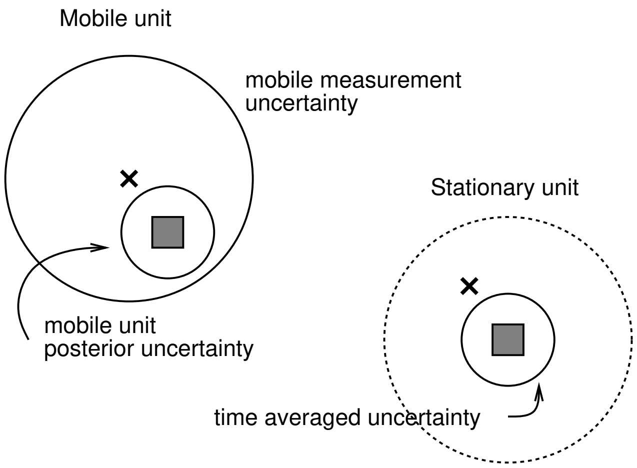


Figure 8: Differential GPS

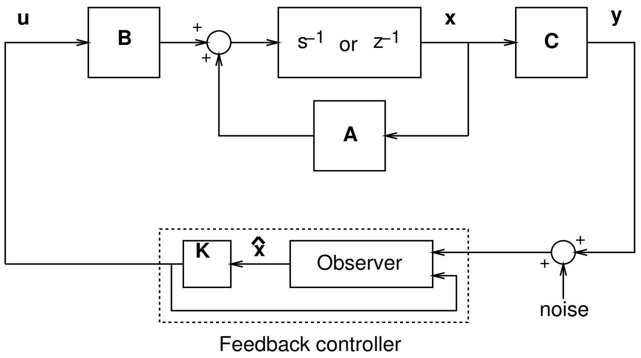


Figure 9: Feedback control using an observer (after Jacobs p217)

1.6 Example: Feedback control

Suppose a system is modelled by state equations (see figure 9):

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \\ \text{or} \quad \mathbf{x_i} &= \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \end{aligned}$$

and we seek to control the system with a control law of the form

$$\mathbf{u} = -\mathbf{K}\mathbf{x}$$

in the presence of disturbances and noise on the sensor outputs. Typically our sensor does not measure the state $\mathbf{x}$ directly, instead measuring some function $\mathbf{y}$ of the state. We might try

$$\mathbf{u} = -\mathbf{K}\mathbf{C}^{-1}\mathbf{y}$$

but it will often be the case that $\mathbf{C}$ is rank-deficient.

Idea: devise an estimator for the state $\mathbf{x}$, $\hat{\mathbf{x}}$, given prior information about $\mathbf{x}$, and a set of observations $\mathbf{y}$, and use the control law

$$\mathbf{u} = -\mathbf{K}\hat{\mathbf{x}}$$

In the context of feedback control such an estimator is known as an **observer**, and a particularly useful and widespread one is the **Kalman filter** which will be discussed in the second part of the course.

2 Probabilistic Tools

We assume a certain amount of basic knowledge from prelims, but not much.

2.1 Independence

Two events A and B are independent if they do not affect each others outcome. Mathematically this is formalized as

$$P(A, B) = P(A)P(B)$$

If the events are not independent then this equation becomes relation

$$P(A, B) = P(A|B)P(B)$$

where the first term on the right hand side means the probability of event A given event B.

2.2 Random Variables

Informally a random variable is a variable which takes on values (either discrete or continuous) at random. It can be thought of as a function of the outcomes of a random experiment.

The probability that a continuous random variable takes on specific values is given by the (cumulative) **probability distribution**:

$$F_X(x) = P(X \leq x)$$

or by the **probability density** function:

$$p_X(x) = \frac{d}{dx} F_X(x)$$

i.e., infinitesimally

$$P(x < X \leq x + dx) = p_X(x)dx$$

From the definition we have the following properties:

$$\int_{-\infty}^{x} p_X(x)dx = F_X(x)$$
$$\int_{-\infty}^{\infty} p_X(x)dx = 1$$
$$\int_{a}^{b} p_X(x)dx = F_X(b) - F_X(a) = P(a < X \leq b)$$

2.3 Expectation and Moments

In many trials, with outcome taking the value x_i with probability p_i, we expect the "average" to be $\sum_i p_i x_i$, hence the **expected value** of a random experiment with n discrete outcomes is given by

$$\text{Expected value of } X = E[X] = \sum_{i=1}^{n} p_i x_i$$

For a continuous random variable we have

$$E[X] = \int_{-\infty}^{\infty} x p_X(x) dx$$

[Exercise: verify that E is a linear operator]

The k**th moment** of X is defined to be $E[X^k]$, i.e.:

$$E[X^k] = \int_{-\infty}^{\infty} x^k p_X(x) dx$$

The first moment, $E[X]$ is commonly known as the **mean**. The second moment is also of particular interest since the **variance** of a random variable can be defined in terms of its first and second moments:

$$\text{Variance of } X = \text{Var}[X] \quad = \quad E[(X - E[X])^2] = E[X^2] - E[X]^2$$

[Exercise: Proof]

2.4 Characteristic Function

The **characteristic function** of a random variable is defined as

$$\Psi_X(\omega) = E[\exp(j\omega X)] = \int_{-\infty}^{\infty} e^{j\omega x} p_X(x) dx$$

Hence

- $E[X^k] = j^n \frac{d^n \Psi_X(\omega)}{d\omega^n}|_{\omega=0}$
- $\Psi_X(\omega)$ is in the form of a Fourier transform, hence
- $p_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-j\omega x} \Psi_X(\omega) d\omega$

2.5 Univariate Normal Distribution

The most important example of a continuous density/distribution in the **normal** or **gaussian** distribution, which is described by the density function (pictured in figure 10)

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2})$$

or by the characteristic function

$$\Psi_X(\omega) = \exp(\mu j\omega - \frac{1}{2}\sigma^2\omega^2)$$

For convenience of notation we write

$$X \sim N(\mu, \sigma^2)$$

to indicate that the random variable X is normally distributed with parameters μ and σ^2, respectively the mean and variance.

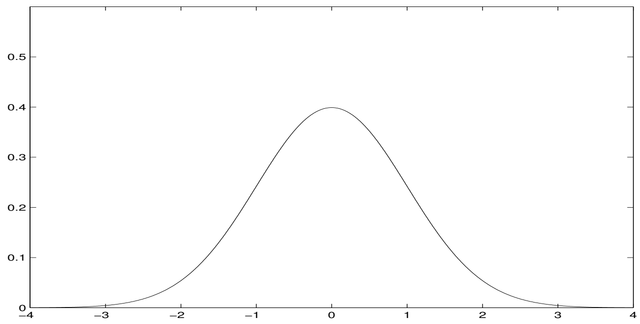


Figure 10: Normal distribution for $\mu = 0, \sigma^2 = 1$

The normal distribution is very important for a number of reasons which will return to later...

2.6 Multiple Random Variables

We begin the discuss of multiple random variables with a revision of material on discrete events.

The **joint probability distribution** of random variables A and B is given by

$$P(A, B) = \text{Probability of events A and B both occurring}$$

Of course this must satisfy

$$\sum_{i,j} P(A_i, B_j) = 1$$

The **conditional probability distribution of A given B** is given by

$$P(A|B) = P(A, B)/P(B)$$

The **marginal distribution** of A is the unconditional distribution of A (similarly for B)

$$P(A) = \sum_j P(A, B_j) = \sum_j P(A|B_j)P(B_j)$$

Combining expressions for $P(A, B)$ and $P(B, A)$ (which are of course equal) we obtain **Bayes' rule**

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}$$

2.7 Discrete example

An example (albeit not a very realistic one) will illustrate these concepts. Consider a digital thermocouple used as an ice-warning device. We model the system by two random variables

$$\begin{aligned} X, &\quad \text{the actual temperature}, 0 \le X \le 2^oC \\ Z, &\quad \text{the sensor reading}, 0 \le Z \le 2^oC) \end{aligned}$$

The joint distribution is given in the "spreadsheet" of $P(X = i, Z = j)$:

	$X = 0$	$X = 1$	$X = 2$	row sum
$Z = 0$	0.32	0.03	0.01	
$Z = 1$	0.06	0.24	0.02	
$Z = 2$	0.02	0.03	0.27	
col sum				

2.8 Multiple Continuous Random Variables

We now consider the continuous equivalents...

The **joint probability distribution** (cumulative) for random variables X and Y is given by

$$F_{XY}(x, y) = P(X \le x, Y \le y)$$

The **joint density** function is defined as

$$p_{XY}(x, y) = \frac{\partial^2 F_{XY}(x, y)}{\partial x \partial y}$$

Hence we have:

$$P(X, Y \in R) = \int\int_R p_{XY}(x, y) dx dy$$

and if R is a differential region then

$$P(x < X \le x + dx, y < Y \le y + dy) = p_{XY}(x, y) dx dy$$

The **marginal** or unconditional density of X (or similarly Y) is given by

$$p_X(x) = \int_{-\infty}^{\infty} p_{XY}(x, y) dy$$

The **conditional** density can be determined by considering the probability that X is in a differential strip, given that Y is in a differential strip:

$$P(x < X \le x + dx | y < Y \le y + dy) = \frac{p_{XY}(x, y) dx dy}{p_Y(y) dy}$$

Now we wave our hands, cancel the dys, and write

$$P(x < X \le x + dx | Y = y) = \frac{p_{XY}(x, y) dx}{p_Y(y)}$$

Hence

$$p(x|y) = p_{X|Y}(x) = \frac{p_{XY}(x,y)}{p_Y(y)}$$

(which is what we would have hoped for or expected). The **conditional moments** of a distribution are given by

$$E[X|Y] = \int_{-\infty}^{\infty} x p_{X|Y}(x) dx$$

Bayes' rule follows directly from the definitions of conditional and joint densities:

$$p_{X|Y}(x) = \frac{p_{Y|X}(y) p_X(x)}{p_Y(y)}$$

Continuous random variables X and Y are **independent** if and only if

$$p_{XY}(x,y) = p_X(x) p_Y(y)$$

2.9 Covariance and Correlation

Th expectation of the product of two random variables is an important quantity. It is given by

$$E[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy p_{XY}(x,y) dx dy$$

If X and Y are independent then $E[XY] = E[X]E[Y]$. Note that this is a necessary *but not sufficient* condition for independence.

The **covariance** of random variables X and Y is defined as

$$\text{Cov}[X,Y] = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

[Exercise: Proof]

The **correlation coefficient**, ρ_{XY}, between two random variables is a normalized measure of how well correlated two random variables are.

$$\rho_{XY} = \frac{\text{Cov}[X,Y]}{\sqrt{\text{Var}[X]}\sqrt{\text{Var}[Y]}}$$

For perfectly correlated variables (i.e. $X = \pm Y$), $\rho_{XY} = \pm 1$, and for completely uncorrelated variables $\rho_{XY} = 0$.

2.10 Continuous Example

Now reconsider the thermocouple example, except now with the (ever so) slightly more realistic assumption that X, the temperature and Z, the sensor output, are now continuously variable between 0 and $2^o C$.

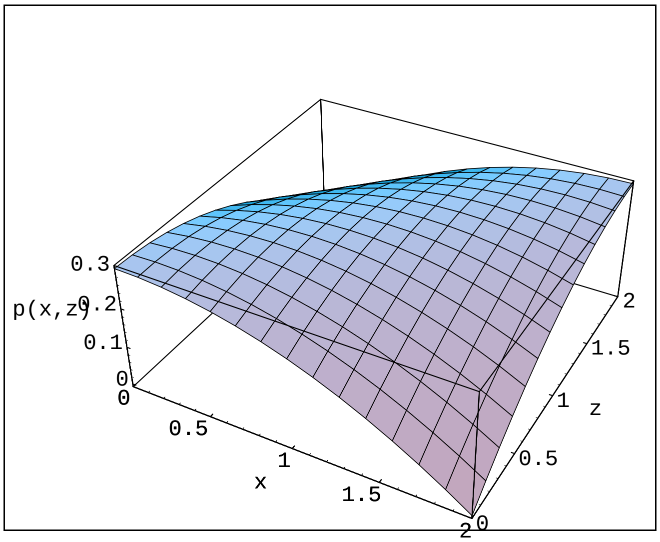


Figure 11: Joint density

Suppose the joint density (illustrated in figure 11) is given by

$$p_{XZ}(x,z) = \frac{3}{10}(1 - \frac{1}{4}(x-z)^2)$$

Marginal distribution:

$$p(x) = \int_0^2 p(x,z)dz =$$

[Exercise: check normalization]

Mean

$$\mu_X = E[X] = \int_0^2 dx\, xp(x) = 1, \quad \mu_Z = 1$$

Variance

$$\sigma_X^2 = \text{Var}[X] = E[(X-\mu_X)^2] =$$

Conditional density

$$p(x|z) = p(x,z)/p(z) =$$

Note that this is simply an x-*slice* through $p(x,z)$, normalized so that $\int_0^2 p(x|z)\,dx = 1$. Also note that by symmetry in x and z in the density function, the marginals $p_X(x)$ and $p_Z(z)$ are identically distributed, as are the conditional densities, $p(x|z)$ and $p(z|x)$. Figure 12 shows the density for $p(z|x)$. Since it is only weakly peaked, we can conclude that it is not a very informative sensor.

Covariance

$$\text{Cov}[X,Z] =$$

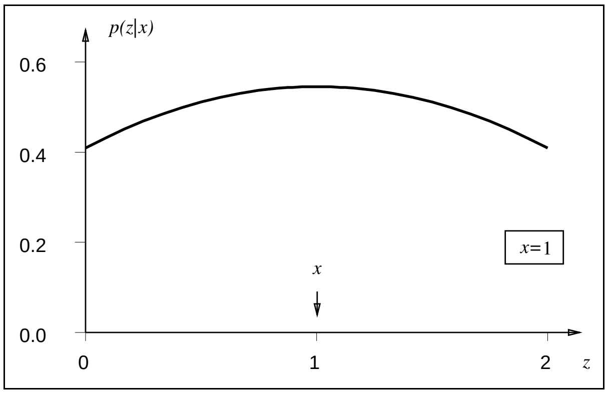


Figure 12: Conditional density $p(z|x)$

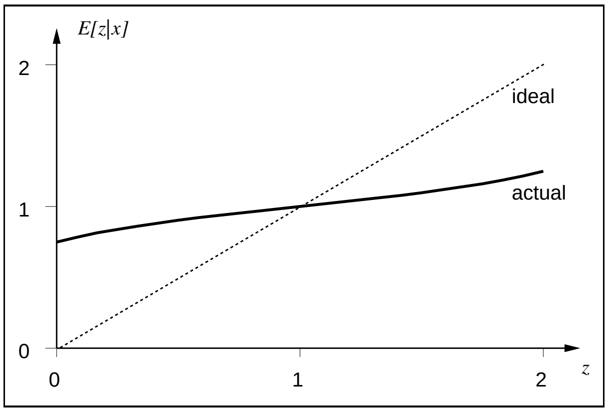


Figure 13: Conditional expectation $E[Z|X]$

Correlation coefficient

$$\rho_{XZ} = \frac{\mathrm{Cov}[X, Z]}{\sigma_x \sigma_z} =$$

Conditional moments

$$E[Z|X] = \int_0^2 z p(z|x) dz$$

Figure 13 shows the the conditional expectation $p(z|x)$ (and also the ideal response), indicating a strong bias in the sensor for all values other than at exactly 1^oC.

2.11 Sum of Independent Random Variables

Given independent random variables, X and Y, with probability densities respectively $p_X(x)$ and $p_Y(y)$, we can derive the density of their sum $Z = X + Y$ by considering the infinitesimal band

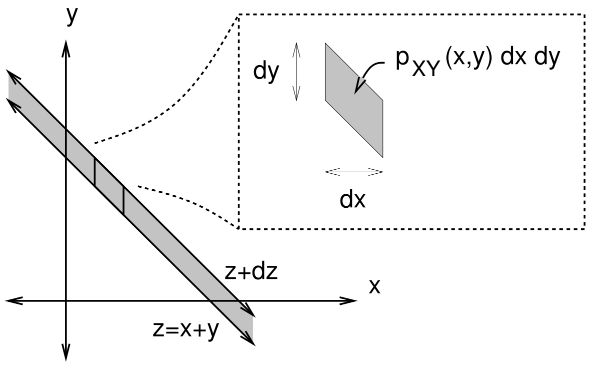


Figure 14: Sum of independent random variables

shown in figure 14, and arguing as follows.

$$\begin{aligned}
p_Z(z)dz &= P(z < Z \leq z + dz) \\
&= \int_{x=-\infty}^{\infty} p_{XY}(x, y) dx\, dy \\
&= \int_{x=-\infty}^{\infty} p_X(x) p_Y(y) dx\, dy \quad \text{X,Y independent} \\
&= \int_{x=-\infty}^{\infty} p_X(x) p_Y(z - x) dx\, dz \quad \text{change of variables}, y = z - x
\end{aligned}$$

Notice that this is a convolution of the two densities. This can either be evaluated "long-hand", or by taking Fourier transforms. It follows immediately from this and the definition of the characteristic function, that the characteristic function of the sum of two random variables is the product of their characteristic functions.

2.12 Properties of the Normal Distribution

We are now in a position to return to the promise of further discussion of the various normal distribution properties which make it both mathematically tractable and so widespread naturally.

Closure under summation The sum of two independent normal distributions is normal. This follows immediately from above, since the product of two normal characteristic functions gives:

$$\exp(\mu_1 j\omega - \frac{1}{2}\sigma_1^2\omega^2). \exp(\mu_2 j\omega - \frac{1}{2}\sigma_2^2\omega^2) = \exp((\mu_1 + \mu_2) j\omega - \frac{1}{2}(\sigma_1^2 + \sigma_2^2)\omega^2)$$

which corresponds to a normal distribution with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$.

We could alternatively do the convolution directly, or using Fourier transforms.

Central limit theorem It is a quite remarkable result, that the distribution of the sum of n independent random variables of any distribution, tends to the normal distribution for n large.

$$\begin{aligned}
&\text{if} \quad X_i \sim D(\mu_i, \sigma_i^2), \quad i = 1, \dots n \\
&\text{then} \quad \sum_i X_i \sim N(\mu, \sigma^2), \quad \text{where } \mu = \sum_i \mu_i, \;\; \sigma_2 = \sum_i \sigma_i^2
\end{aligned}$$

```
function clt(n)
 % demonstrate convergence towards normal
 % for multiple i.i.d uniform random variables

 t = [-10:0.01:10]; % define U(0,1)
 pulse = zeros(2001,1);
 pulse(900:1100) = ones(201,1);
 plot(t, pulse);    % plot U(0,1)
 hold on;
 pause;

 res = pulse;
 for i=1:n-1        % now iterate convolution
     new = conv(pulse,res);
     res = new(1000:3000)*0.005;
     plot(t, res);
     pause;
 end
```

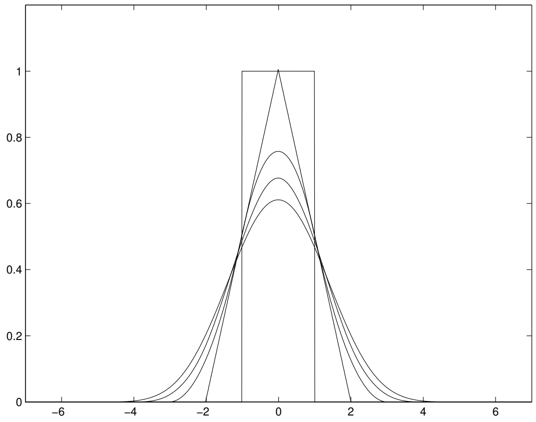


Figure 15: Matlab code and resulting plots showing the distribution of a sum of n uniform random variables

A general proof can be found in Riley. Figure 15 demonstrates the result for the case of i.i.d random variables with uniform distribution.

Closure under pooling Under suitable assumptions (independence, uniform prior), if

$$p(x|z_i) \sim N(\mu_i, \sigma_i), \quad i = 1, 2$$

then

$$p(x|z_1, z_2) \sim N(\mu, \sigma)$$

2.13 Multivariate Random Variables

Consider n random variables $X_1, X_2, \ldots X_n$, and let

$$\mathbf{X} = [X_1, \ldots X_n]^\top$$

Now we define:

$$E[\mathbf{X}] = [E[X_1], \ldots E[X_n]]^\top$$

$$\text{Cov}[\mathbf{X}] = E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^\top]$$

From the definition of the **covariance matrix** we can see that the ijth entry in the matrix is

$$\text{Cov}(\mathbf{X})|_{ij} = \begin{cases} \text{Var}[X_i] & \text{if } i = j \\ \text{Cov}[X_i, Y_j] & \text{if } i \neq j \end{cases}$$

Clearly the covariance matrix must be symmetric.

[Exercise: prove that the covariance matrix satisfies $\mathbf{x}^\top \mathbf{C} \mathbf{x} \geq 0, \forall \mathbf{x}$ – i.e. that $\mathbf{C}$ is positive semi-definite]

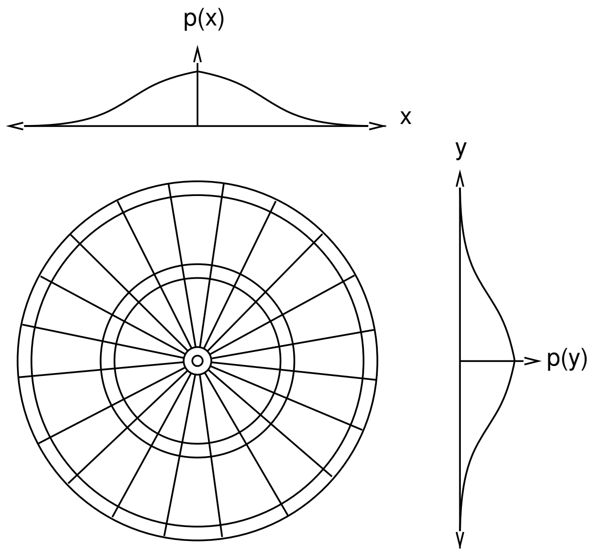


Figure 16: The centre of the dart board is the cartesian coordinate (0,0)

2.14 Multivariate Normal

$\mathbf{X}$ is a vector of random variables, with corresponding mean μ and covariance matrix $\mathbf{C}$.

The random variables which comprise $\mathbf{X}$ are said to be **jointly normal** if their joint density function is given by

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\mathbf{C}|^{1/2}} \exp(-\frac{1}{2}(\mathbf{x}-\mu)^\top \mathbf{C}^{-1}(\mathbf{x}-\mu))$$

Note that the exponent $\mathbf{x} - \mu)^\top \mathbf{C}^{-1}(\mathbf{x} - \mu)$ is a scalar, and that the covariance matrix must be non-singular for this to exist.

When $n = 1$ the formula clearly reduces to the univariate case. Now consider the bi-variate case, $n = 2$, beginning with an example. A dart player throws a dart at the board, aiming for the bulls-eye. We can model where the dart is expected to hit the board with two normally distributed random variables, one for the x-coordinate and one for the y-coordinate (see figure 16); i.e.

$$X \sim N(0, \sigma_X^2), \quad Y \sim N(0, \sigma_Y^2)$$

Since X and Y are independent, we have that $P_{XY}(x, y) = p_X(x)p_Y(y)$, so

$$\begin{aligned}
p_{XY}(x,y) &= \frac{1}{\sqrt{2\pi}\sigma_X} \exp(-\frac{x^2}{2\sigma_X^2}).\frac{1}{\sqrt{2\pi}\sigma_Y} \exp(-\frac{y^2}{2\sigma_Y^2}) \\
&= \frac{1}{2\pi\sigma_X\sigma_Y} \exp(-\frac{1}{2}(\frac{x^2}{\sigma_X^2} + \frac{y^2}{\sigma_Y^2})) \\
&= \frac{\sqrt{|\mathbf{S}|}}{2\pi} \exp(-\frac{1}{2}\mathbf{x}^\top \mathbf{S}\mathbf{x})
\end{aligned}$$

where $\mathbf{x} = \begin{bmatrix} x & y \end{bmatrix}^\top$ and $\mathbf{S} = ??$

Plots of the surface p_{XY} are shown in figures 17, and 18 for

$$\mu = \begin{bmatrix} 0 & 0 \end{bmatrix}^\top, \quad \mathbf{C} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Note that the "iso-probability" lines

$$\mathbf{x}^\top \mathbf{S} \mathbf{x} = d^2$$

are circles.

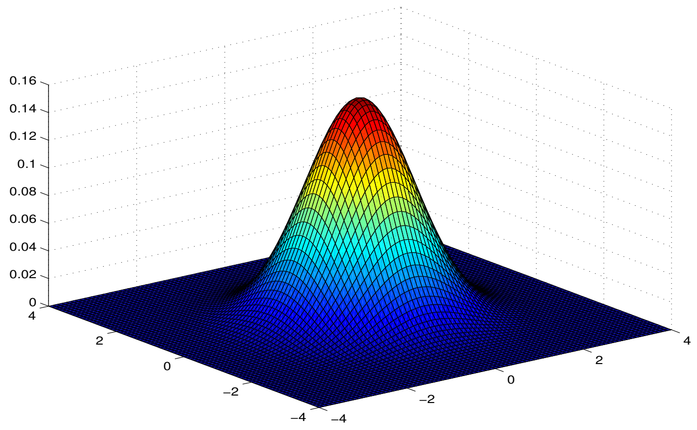


Figure 17: A bivariate normal, surface plot

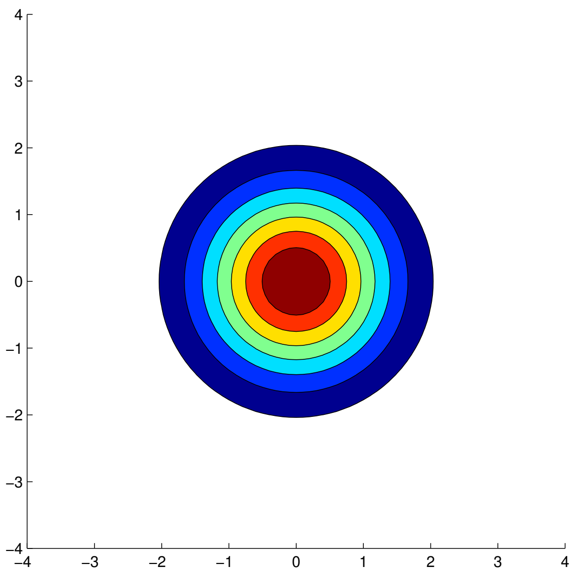


Figure 18: A bivariate normal, contour plot

Now suppose that the random variables X and Y are not independent, but instead have correlation coefficient $\rho \neq 0$. The covariance matrix is then:

$$\mathbf{C} = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$$

We can see even without plotting values, that the iso-probability contours will be elliptical:

$$\mathbf{x}^\top \mathbf{S}\mathbf{x} = \mathbf{x}^\top \begin{bmatrix} \frac{1}{(1-\rho^2)\sigma_X^2} & \frac{-rho}{(1-\rho^2)\sigma_X\sigma_Y} \\ \frac{-rho}{(1-\rho^2)\sigma_X\sigma_Y} & \frac{1}{(1-\rho^2)\sigma_Y^2} \end{bmatrix} \mathbf{x} = d^2$$

Plots are shown in figures in figures 19, and 20 for

$$\mu = \begin{bmatrix} 0 & 0 \end{bmatrix}^\top, \quad \mathbf{C} = \begin{bmatrix} 1 & 1.2 \\ 1.2 & 4 \end{bmatrix}$$

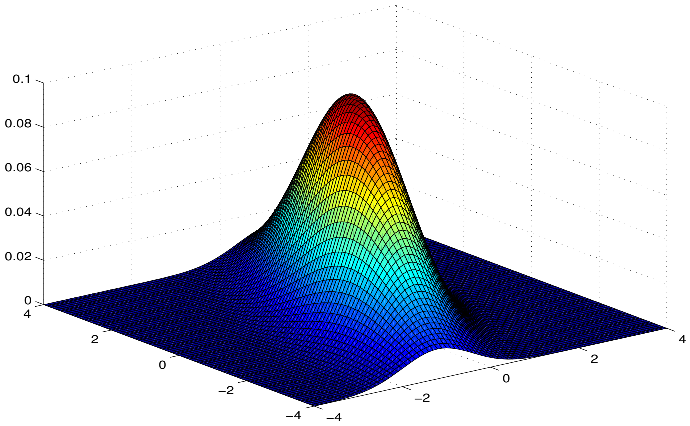


Figure 19: A bivariate normal, surface plot

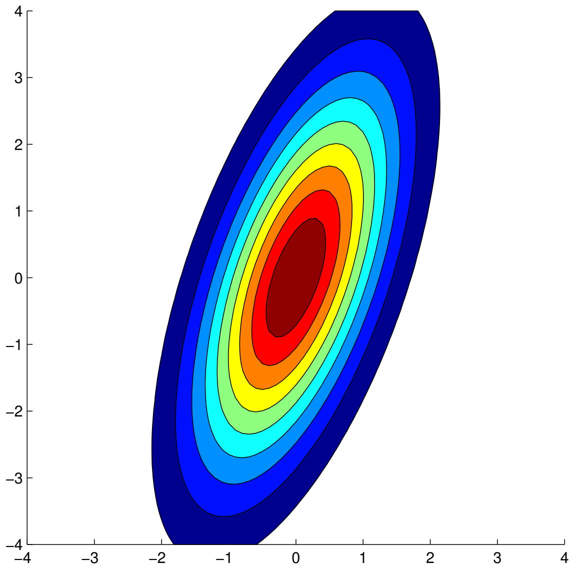


Figure 20: A bivariate normal, contour plot

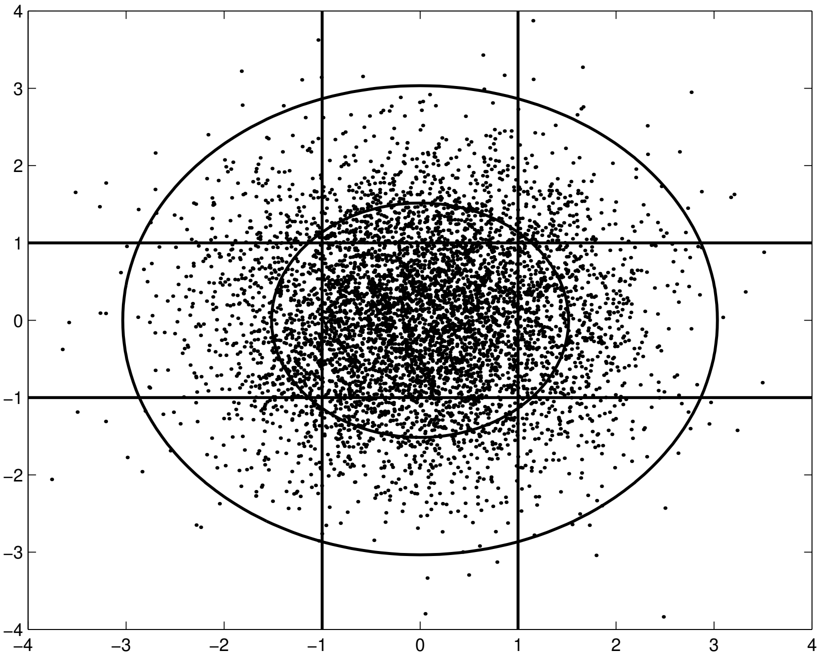


Figure 21: Confidence regions and intervals

2.15 Confidence Intervals and Regions

A confidence region (or interval) R around $\hat{\mathbf{x}}$ at confidence level $p \leq 1$ is a region defined such that

$$P(\mathbf{x} \in R) = p$$

We can use it to say (for example) "there is a 99% chance that the true value falls within a region around the measured value", or that "there is only a 5% chance that this measurement is not an outlier".

Typically choose a regular shape such as an ellipse. In figure 21 a random experiment has been conducted with observations generated from a bivariate normal distribution. Various confidence regions have been superimposed.

Relating variance/covariance and confidence

For a univariate random variable of any distribution, the **chebychev inequality** relates the variance to a confidence, by:

$$P(|X - \mu| \geq d) \leq \frac{\text{Var}[X]}{d^2}$$

This however gives quite a weak bound. For a univariate normal distribution $N(\mu, \sigma_2$, the bound is *much* tighter:

$$\begin{aligned} P(|X - \mu| \leq \sigma) &\approx 0.67 \\ P(|X - \mu| \leq 2\sigma) &\approx 0.95 \\ P(|X - \mu| \leq 3\sigma) &\approx 0.997 \end{aligned}$$

s can be verified by looking up standard normal distribution tables.

If $\mathbf{x}$ of dimension n is normally distributed with mean zero and covariance $\mathbf{C}$, then the quantity $\mathbf{x}^\top \mathbf{C}^{-1}\mathbf{x}$ is distributed as a χ^2-distribution on n degrees of freedom.

The region enclosed by the ellipse

$$\mathbf{x}^\top \mathbf{C}^{-1}\mathbf{x} = d^2$$

defines a confidence region

$$P(\chi^2_\nu < d^2) = p$$

where p can be obtained from standard tables (e.g. HLT) or computed numerically.

2.16 Transformations of Random Variables

A transformation from one set of variables (say, inputs) to another (say, outputs) is a common situation. If the inputs are stochastic (i.e. random) how do we characterize the ouptuts?

Suppose g is an invertible transformation

$$\mathbf{y} = g(\mathbf{x})$$

and we will consider $\mathbf{x}$ to be a realisation of a random variable $\mathbf{X}$. If we know the density functioon for $\mathbf{X}$ is given by $p_{\mathbf{X}}(\mathbf{x})$ then $\mathbf{y}$ is a realisation of a random variable $\mathbf{Y}$ whose density is given by

$$p_{\mathbf{Y}}(\mathbf{y}) = p_{\mathbf{X}}(h(\mathbf{y}))|Jh(\mathbf{y})|$$

where

$$h = g^{-1}, \quad \mathbf{x} = h(\mathbf{y})$$

and $|Jh(\mathbf{y})|$ is the Jacobian determinant of h.

Proof:

Example: rectangular to polar coordinates

Suppose we have a bivariate joint density function

$$p_{XY}(x, y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$$

i.e. bivariate normal density with X and Y independent with zero mean and identical variance.

Now we wish to find the corresponding density function in terms of polar coordinates r and θ (see figure 22):

$$\begin{aligned} x &= r\cos\theta \\ y &= r\sin\theta \end{aligned}$$

2.17 Linear Transformation of Normals

As we have seen a normal distribution can be completely specified by its mean and variance (1st and 2nd moments). Furthermore, the linear combinations of normals are also normal. Suppose we effect the transformation

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{c}$$

where $\mathbf{A}$ and $\mathbf{c}$ are a matrix and vector constant (repectively), and $\mathbf{x}$ is a realisation of a random variable $\mathbf{X}$ which is normally distributed with mean $\boldsymbol{\mu}$ and covariance $\mathbf{C}$.

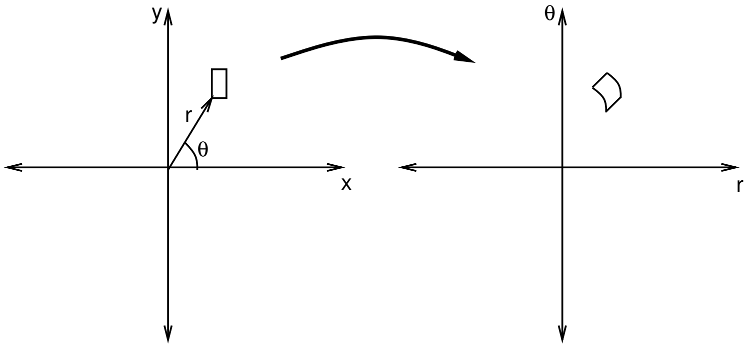


Figure 22: Transformation of a random variable

We now have that

$$\begin{aligned} E[\mathbf{Y}] &= \mathbf{A}E[\mathbf{X}] + \mathbf{c} \\ \mathrm{Cov}[\mathbf{Y}] &= \mathbf{A}\mathrm{Cov}[\mathbf{X}]\mathbf{A}^\top \end{aligned}$$

[Exercise: Proof]

The **information matrix** is the inverse of the covariance matrix. Suppose we choose a linear transformation which diagonalizes $\mathbf{S} = \mathbf{C}^{-1}$. Since $\mathbf{S}$ is symmetric, positive semi-definite, its eigendecomposition is real and orthonormal:

$$\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$$

Let us transform $\mathbf{x}$ according to

$$\mathbf{y} = \mathbf{V}^\top(\mathbf{x} - \boldsymbol{\mu})$$

then the random variables $\mathbf{Y}$ are normally distributed with mean zero and diagonal covariance (i.e. the individual random variables are independent [proof?]). Notice that the eigenvectors of $\mathbf{S}$ (equivalently $\mathbf{C}$) are the principal axes of the ellipsoids of constant probability density – i.e. the confidence region boundaries. The square roots of the eigenvalues give the relative sizes of the axes. A small eigenvalue of $\mathbf{S}$ (equivalently a large eigenvalue of $\mathbf{C}$) indicates a lack of information (high degree of uncertainty), while a large eigenvalue of $\mathbf{S}$ (equivalently a small eigenvalue of $\mathbf{C}$), indicates low degree of uncertainty.

A further transformation $\mathbf{w} = \mathbf{\Lambda}^{1/2}\mathbf{y}$ yields a vector $\mathbf{w}$ of **standard normals**, $w_i \sim N(0, 1)$.

3 Estimators

We consider the problem of estimating a parameter x based on a number of observations z_i in the presence of noise on the observations w_i (i.e. we will treat the w_i as zero mean random variables)

$$z_i = h(i, x, w_i)$$

To begin, we will look briefly at general important properties of estimators.

3.1 Mean square error

The **mean square error** of an estimator $\hat{\mathbf{x}}$ of a parameter $\mathbf{x}$ is defined to be

$$MSE_{\mathbf{x}}(\hat{\mathbf{x}}) = E[(\hat{\mathbf{x}} - \mathbf{x})^2]$$

Note that here we are potentially dealing with vector quantities and so $(\hat{\mathbf{x}} - \mathbf{x})^2$ has the obvious meaning

$$(\hat{\mathbf{x}} - \mathbf{x})^2 = ||\hat{\mathbf{x}} - \mathbf{x}||^2 = (\hat{\mathbf{x}} - \mathbf{x})^\top(\hat{\mathbf{x}} - \mathbf{x})$$

3.2 Bias

An estimator $\hat{\mathbf{x}}$ of a parameter $\mathbf{x}$ is said to be **unbiased** if (and only if)

$$E[\hat{\mathbf{x}}] = \mathbf{x}$$

i.e. its distribution is "centred" on the true value.

The mean square error then becomes equal to the variance of the estimator plus the square of the bias. Oviously, therefore, an unbiased estimator has mean square error equal to its variance. For the most part we will consider only unbiased estimators.

Example: sample mean

Given a set of independent measurements $z_i, i = 1, \ldots n$, where

$$z_i = \mu + w_i, \quad w_i \sim N(0, \sigma^2)$$

(hence $z_i \sim N(\mu, \sigma^2)$)

We can estimate the mean of the distribution using the sample mean, given by $\bar{x} = 1/n \sum_i z_i$. It is clearly unbiased since

$$E[\bar{x}] \quad = \quad E[\frac{1}{n}\sum_i z_i] = \frac{1}{n}\sum_i E[z_i] = \frac{1}{n} n\mu = \mu$$

The variance of the estimator is

$$E[(\bar{x} - \mu)^2] \quad = \quad E[(\sum (z_i - \mu))^2]/n^2 = \sigma^2/n$$

Example: sample variance

Consider estimating the variance using the estimator

$$\frac{1}{n}\sum_i (z_i - \bar{x})^2$$

This estimate is biased!

Proof:

For this reason you will often see the sample variance written as

$$\frac{1}{n-1}\sum_i (z_i - \bar{x})^2$$

3.3 Consistency and Efficiency

A mathematical discussion of these is beyond the scope of the course (and what most people require), but – perhaps moreso than unbiasedness – these are desirable properties of estimators.

A **consistent estimator** is one which provides an increasingly accurate estimate of the parameter(s) as n increases. Note that the sample mean is clearly consistent since the variance of the sample mean is σ^2/n, which decreases as $n \to \infty$.

An **efficient estimator** is one which minimizes the estimation error (it attains its theoretical limit)

3.4 Least Squares Estimation

Suppose that we have a set of n observations z_i which we wish to fit to a model with m parameters θ_j, where the model predicts a functional relationship between the observations and the model parameters:

$$z_i = h(i; \theta_1, \ldots \theta_m) + w_i$$

In the absence of any statistical knowledge about the noise w_i and the model parameters we might consider a minimizing a least-squares error:

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\min_{\theta_1 \ldots \theta_m} \sum_{i=1}^{n} w_i^2 \\
&= \arg\min_{\theta_1 \ldots \theta_m} \sum_{i=1}^{n} (z_i - h(i; \theta_1, \ldots \theta_m))^2
\end{aligned}$$

Example 1: Sample mean (again)

Suppose

$$z_i = \theta + w_i$$

(i.e. $h(i;\theta) = \theta$) So the least squares estimate is given by

$$\hat{\theta} = \arg\min_{\theta} \sum_i (z_i - \theta)^2$$

$$\text{hence} \quad \hat{\theta} = 1/n \sum_i z_i$$

i.e. the sample mean.

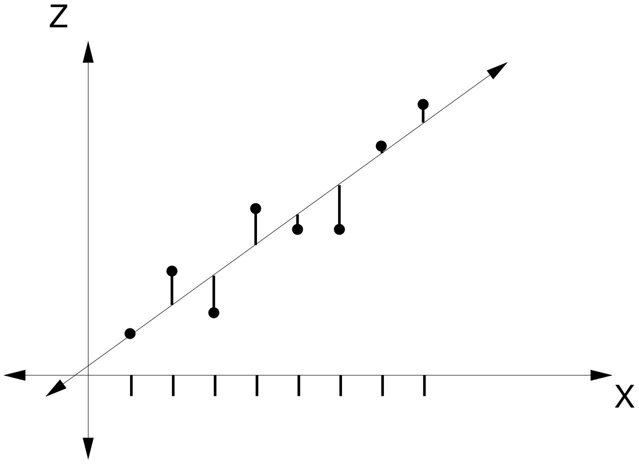


Figure 23: Fitting data to a straight line

Example 2: Fitting data to a straight line

Suppose

$$z_i = \alpha x_i + \beta + w_i$$

(see figure 23) where noisy observations z_i are made at known locations x_i (assumed noise free). The least squares estimate is given by

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_i (z_i - \alpha x_i - \beta)^2, \quad \boldsymbol{\theta} = [\alpha \;\; \beta]^\top$$

General Linear Least Squares

In the examples above, $h(i; \theta_1, \ldots \theta_m)$ is linear in the parameters θ_j. Recall that you have encountered a more general form than either of these examples in your second year Engineering Computation; i.e. polynomial regression, where the model was a linear combination of the monomials x^j.

$$z_i = \sum_{j=0}^{m} \theta_j (x_i)^j$$

These are all cases of a general formulation for **linear least squares**, in which the model is a linear combination of a set of arbitrary **basis functions** which may be wildly non-linear – it is the dependence on θ_j which must be linear. The model expressed as

$$\mathbf{z} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

(so for polynomial regression $\mathbf{H}$ is the so-called Vandermonde matrix). The sum of squares then becomes

$$J = \sum_i w_i^2 = \mathbf{w}^\top \mathbf{w} = (\mathbf{z} - \mathbf{H}\boldsymbol{\theta})^\top (\mathbf{z} - \mathbf{H}\boldsymbol{\theta})$$

Differentiating J w.r.t. $\boldsymbol{\theta}$ and setting equal to zero yields

$$\mathbf{H}^\top \mathbf{z} = \mathbf{H}^\top \mathbf{H}\boldsymbol{\theta}$$

which are the **normal equations** for the problem. If rank$\mathbf{H} = m$ then $\mathbf{H}^\top\mathbf{H}$ is invertible and we have a unique least squares estimate for $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\mathbf{z}$$

If rank$\mathbf{H} < m$ then there exists a family of solutions.

Weighted Least Squares

If we have some idea (how?) of the relative reliability of each observation we can weight each individual equation by a factor σ^{-2}, and then minimize:

$$\sum_i \sigma_i^{-2}(z_i - h(i;\boldsymbol{\theta}))^2 = (\mathbf{z} - \mathbf{H}\boldsymbol{\theta})^\top \mathbf{R}^{-1}(\mathbf{z} - \mathbf{H}\boldsymbol{\theta})$$

where

$$\mathbf{R} = \mathrm{diag}(\sigma_1^2, \ldots \sigma_n^2)$$

More generally $\mathbf{R}$ could be *any* $n \times n$ symmetric, positive definite weighting matrix.

The **weighted least squares estimate** is then given by

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{z}$$

Note that these results have no probabilistic interpretation. They were derived from an intuition that minimizing the sum of residuals was probably a good thing to do. Consequently least squares estimators may be preferred to others when there is no basis for assigning probability density functions to $\mathbf{z}$ (i.e. $\mathbf{w}$) and $\boldsymbol{\theta}$.

3.5 Maximum Likelihood Estimation

Alternatively, we can adopt the **maximum likelihood** philosophy, where we take as our estimate $\hat{\boldsymbol{\theta}}$ that value which maximizes the probability that the measurements $\mathbf{z}$ actually occurred, taking into account known statistical properties of the noise on them $\mathbf{w}$.

For the general linear model above the density of z_i given $\boldsymbol{\theta}$, $p(z_i|\boldsymbol{\theta})$ is just the density of w_i centred at $\mathbf{H}\boldsymbol{\theta}$. If the $\mathbf{w}$ is zero mean normal with covariance $\mathbf{R}$ (w_i not necessarily independent, so $\mathbf{R}$ not necessarily diagonal), then

$$p(\mathbf{z}|\boldsymbol{\theta}) = \frac{1}{(2\pi)^{(n/2)}|\mathbf{R}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{z} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{R}^{-1}(\mathbf{z} - \mathbf{H}\boldsymbol{\theta})\right]$$

To maximize this probability we minimize the exponent, which is clearly equivalent to the weighted least squared expression from the previous section, with the weight matrix determined by the covariance matrix $\mathbf{R}$ – i.e. a probabilistic basis for choosing the weights now exists. The result, restated, is

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{z}$$

This estimate is known as the **maximum likelihood estimate** or **MLE**. We have thus shown that for normally distributed measurement errors, the least squares (LS) and maximum likelihood (MLE) estimates coincide.

Example (non-Gaussian)

Let us consider an example which does not involve normal distributions (for a change). Suppose instead, that the lifetime of a part is given by a random variable with **exponential distribution**:

$$p_X(x) = ae^{-ax}, \quad x > 0$$

We now measure the lifetime of a set of such parts $z_1, \ldots, z_n$ in order to estimate the parameter a in the exponential distribution. To do this, we maximize the **likelihood function** $L(a) = p(\mathbf{z}|a)$ over the parameter a:

$$\begin{aligned}
\hat{a}_{MLE} &= \arg\max_a \; L(a) \\
&= \arg\max_a \; \prod_{i=1}^{n} ae^{-az_i} \\
&= \arg\max_a \; n \log a - a \sum z_i \\
&= n / \sum z_i
\end{aligned}$$

If, after some time T some of the parts have not yet failed (say parts $m+1, \ldots n$), then we can measure the probability that they are still working using the cumulative distribution:

$$P(X_i > T) = e^{-aT}$$

Now the probability of making the measurements, given the true value of a is

$$P = \prod_{i=1}^{m} p(z_i|a) dz_i \times \prod_{i=m+1}^{n} e^{-aT}$$

Maximizing this probability yields

$$\begin{aligned}
\hat{a}_{MLE} &= \arg\max_a \; \prod_{i=1}^{m} p(z_i|a) dz_i \times \prod_{i=m+1}^{n} e^{-aT} \\
&= \arg\max_a \; a^m \exp(-a \sum_{i=1}^{m} z_i - (n-m)T \\
&= \arg\max_a \; m \log a - a \sum z_i - a(n-m)T \\
&= \frac{m}{\sum z_i - (n-m)T}
\end{aligned}$$

Note that the dz_i are independent of the parameter a and so can be cancelled.

3.6 Bayes' Inference

If in addition to the noise distribution for the observations, we also have a prior for the parameter(s), we can employ Bayes' rule to obtain a posterior density for the parameter, given the observations:

$$p(x|z) = \frac{p(z|x)p(x)}{p(z)} = \frac{\text{likelihood} \times \text{prior}}{\text{normalizing factor}}$$

Example: Uniform

- Sensor model: $p(z|x)$
- Prior information: $p(x)$
- Posterior from Bayes' rule: $p(x|z) \propto p(z|x)p(x)$

Now recall figure 6, reproduced in figure 24.

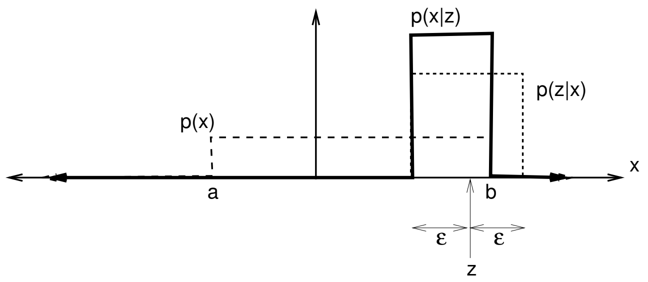


Figure 24: Combining the prior $p(x)$ with a measurement $p(z|x)$ to obtain the posterior

Example: Normal

- Sensor model: $p(z|x) \propto \exp(-(z-x)^2/2\sigma_z^2)$
- Prior information: $p(x) \propto \exp(-(x-\mu_x)^2/2\sigma_x^2)$
- Posterior from Bayes' rule:

$$p(x|z) \propto p(z|x)p(x), \quad X|Z \sim N(\mu, \sigma^2)$$

Hence we have:

- Posterior density: $p(x|z) =$
- Variance: $\sigma^{-2} =$
- Mean: $\mu =$

Example: Independent Sonar Distance Sensors

Suppose we have a point whose likely range centres at 150mm and is normally distributed with standard deviation 30; i.e. $X \sim N(150, 30^2)$ is the prior.

We have two (unbiased) sonar sensors giving independent measurements of the point's range:

$$Z_1|X \sim N(x, 10^2), \quad Z_2|X \sim N(x, 20^2)$$

What can we say about the posterior when:

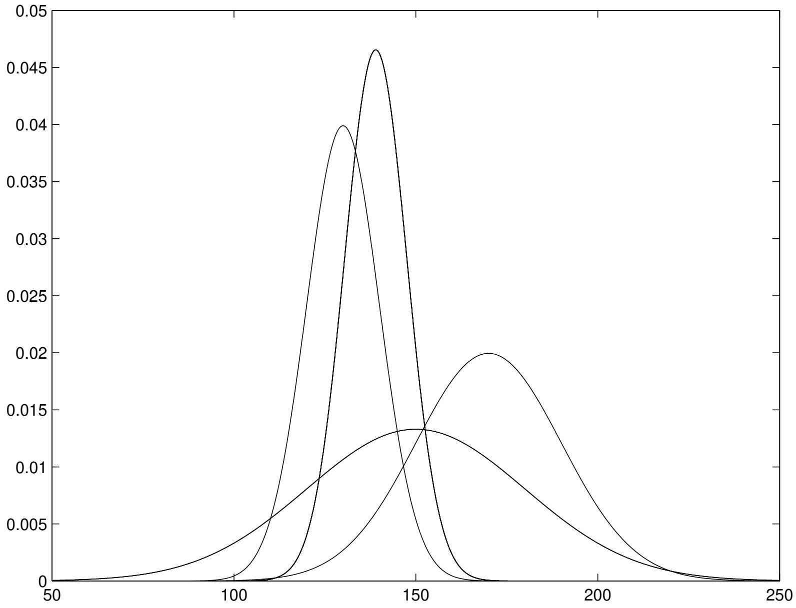


Figure 25: Pooling information from sensor readings and prior information

1. prior of X as above and sensor reading $z_1 = 130$
2. prior of X as above and sensor reading $z_1 = 250$
3. no prior, sensor readings $z_1 = 130, z_2 = 170$
4. prior of X as above and sensor readings $z_1 = 130, z_2 = 170$ (see figure 25)
5. biased sensor $Z_1|X \sim N(x + 5, 10)$

3.7 MMSE and MAP Estimation

In the previous sections we were computing, using Bayes' rule, the posterior density of a parameter (or parameters) given observed data and a prior model for the parameter.

Here, we decide to choose some representative point from the posterior as an estimate of our parameter. Two "reasonable" ideas present themselves:

1. Choose the estimate to be the mean of the posterior $E[x|z]$, which is what would would get by requiring our estimate to minimize the mean square error. Such an estimate is known (funnily enough) as a **minimum mean square error (MMSE) estimate**. Furthermore if the estimate is unbiased then it is a **minimum variance (unbiased) estimate (MVUE)**.

$$\hat{x}_{MMSE} = \arg\min_{\hat{x}} E[(\hat{x} - x)^2|z] = E[x|z]$$

2. Choose the estimate as the mode, or peak value, of the posterior distribution. This is named **maximum a posteriori estimation** (or **MAP**). We obtain the estimate by maximizing $p(x|z)$ over x. Note that the denominator of $p(x|z)$ does not depend on the parameter x and therefore does not affect the optimization – for this reason we often omit it:

$$p(x|z) \propto \text{likelihood} \times \text{prior}$$

and

$$\hat{x}_{MAP} = \arg\max_x \; p(x|z) = \arg\max_x \; p(z|x)p(x)$$

Clearly if $x|z$ (the posterior density) is normal then the MAP estimate $\hat{x}_{MAP}$ is the mean $E[x|z]$, since the normal density attains its maximum value at the mean.

3.8 Multivariate Pooling

- Sensor model:

$$p(\mathbf{z}|\mathbf{x}) \propto \exp\left(-\frac{1}{2}(\mathbf{z}-\mathbf{H}\mathbf{x})^\top \mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}\mathbf{x})\right)$$

- Prior information:

$$p(\mathbf{x}) \propto \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{x_{0x}})^\top \mathbf{P_0}^{-1}(\mathbf{x}-\mathbf{x_{0x}})\right)$$

- Posterior from Bayes' rule:

$$p(\mathbf{x}|\mathbf{z}) \propto p(\mathbf{z}|\mathbf{x})p(\mathbf{x}), \quad \mathbf{x}|\mathbf{z} \sim N(\hat{\mathbf{x}}, \mathbf{C})$$

Hence we have:

- Posterior density: $p(\mathbf{x}|\mathbf{z}) =$
- Covariance: $\mathbf{C} =$
- Mean: $\hat{\mathbf{x}} =$

Example 1: Visual Navigation, 2D Measurements

Suppose the bivariate gaussian prior for the (x, y) location of a beacon is known. We have a sensor which gives an unbiased 2D measurement of position (e.g. from binocular cameras) corrupted by zero mean gaussian noise; i.e:

$$\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{w}, \quad \mathbf{H} = \mathbf{I}_2$$

Both are shown in figure 26(left) (using the values indicated in the matlab fragment below).

```
>> x = [-1; -1];  P = [2 1; 1 3];  % prior
>> z = [1; 2];  R = [1 0; 0 1];    % sensor
>>
>> S = inv(P);  Ri = inv(R);
>> Pnew = inv(S + Ri);
>> xnew = Pnew * (S*x + Ri*z);
```

The posterior computed in the matlab fragment above is shown in figure 26(right)

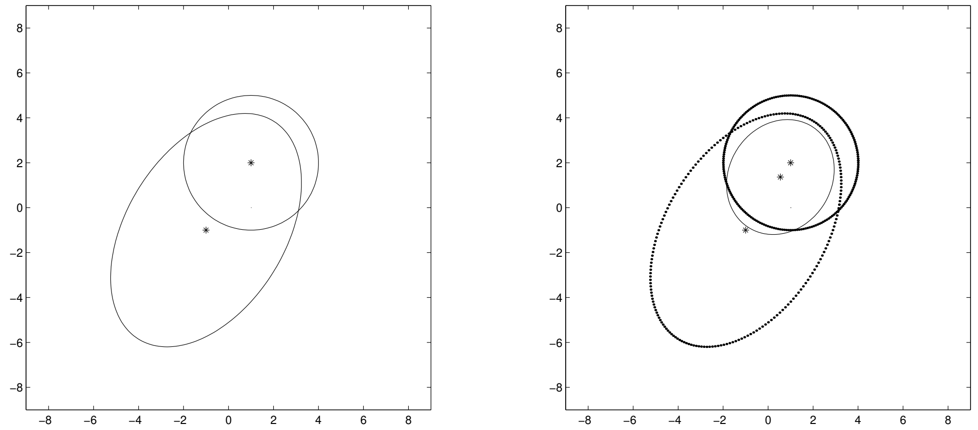


Figure 26: Mean and 3-sigma covariance ellipses for prior and measurement (left), and mean and 3-sigma covariance ellipse for posterior (right)

Example 2: Visual Navigation, 1D Measurements

Now suppose that our visual system is monocular – i.e. we can only make 1D measurements. Further suppose that our system for localizing the beacon takes a measurement z parallel to the y-axis; i.e.

$$z = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + w, \quad \mathbf{H} = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

where $w \sim N(0, \sigma^2)$

```
>> x = [-1; -1];  P = [2 1; 1 3];  % prior
>> z = 1;  R = 2;  H = [1 0];      % sensor
>>
>> Pnew = inv(H'*inv(R)*H + inv(P));
>> xnew = Pnew * (inv(P)*x + H'*inv(R)*z);
```

The prior, measurement and posterior are shown in figure 27.

3.9 Multivariate MAP

More generally, suppose that we have a vector of measurements $\mathbf{z} = [z_1, \ldots z_n]^\top$ corrupted by zero mean, gaussian noise. Then the likelihood function is

$$p(\mathbf{z}|\mathbf{x}) = k \exp\left(-\frac{1}{2}(\mathbf{z} - \mathbf{H}\mathbf{x})^T \mathbf{R}^{-1} (\mathbf{z} - \mathbf{H}\mathbf{x})\right)$$

If the prior on $\mathbf{x}$ is multivariate normal with mean $\mathbf{x_0}$ and covariance $\mathbf{P_0}$ then the posterior has distribution

$$p(\mathbf{x}|\mathbf{z}) = k' \exp\left(-\frac{1}{2}(\mathbf{z} - \mathbf{H}\mathbf{x})^T \mathbf{R}^{-1} (\mathbf{z} - \mathbf{H}\mathbf{x})\right) \times \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x_0})^T \mathbf{P_0}^{-1} (\mathbf{x} - \mathbf{x_0})\right)$$

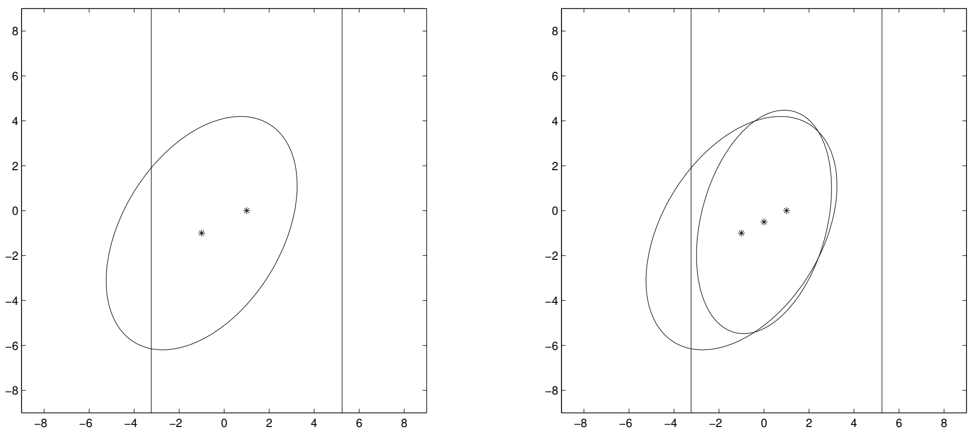


Figure 27: Mean and 3-sigma covariance ellipses for prior and measurement (left), and mean and 3-sigma covariance ellipse for posterior (right)

which is maximized when

$$\frac{\partial}{\partial \mathbf{x}}\left[-\frac{1}{2}(\mathbf{z}-\mathbf{H}\mathbf{x})^T\mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}\mathbf{x})-\frac{1}{2}(\mathbf{x}-\mathbf{x_0})^T\mathbf{P_0}^{-1}(\mathbf{x}-\mathbf{x_0})\right]=0$$

i.e. when

$$\hat{\mathbf{x}}=(\mathbf{P_0}^{-1}+\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H})^{-1}\left(\mathbf{P_0}^{-1}\mathbf{x_0}+\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{z}\right)$$

An alternative derivation of this equation comes if we search for a **minimum variance Bayes' estimate** by minimizing

$$J=\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}(\hat{\mathbf{x}}-\mathbf{x})^\top\mathbf{S}(\hat{\mathbf{x}}-\mathbf{x})p(\mathbf{x}|\mathbf{z})dx_1\dots dx_m$$

where $\mathbf{S}$ is any symmetric, positive definite matrix, and does not affect the result. Differentiate J w.r.t. $\hat{\mathbf{x}}$ and set equal to zero to see that J is minimized when

$$\hat{\mathbf{x}}=E[\mathbf{x}|\mathbf{z}]=(\mathbf{P_0}^{-1}+\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H})^{-1}\left(\mathbf{P_0}^{-1}\mathbf{x_0}+\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{z}\right)$$

Comparing this result to the previous estimators we see that:

- As $\mathbf{P_0}^{-1}\to 0$ (i.e. prior is less and less informative), we obtain the expression for the MLE.
- If the measurement errors are uncorrelated and equal variance then we obtain the expression for unweighted LS.
- The mean and the mode (peak) of a gaussian coincide (in this case the gaussian is the multivariate one for $\mathbf{x}$ conditioned on the measurements $\mathbf{z}$), therefore the minimum variance estimate and the MAP estimate are the same as we have probably anticipated from the similar univariate result above.
- This equation corresponds exactly to a Kalman Filter update of the state mean.

3.10 Recursive regression

In our previous formulations of our estimators we have for the most part assumed that data were available to be processed in "batch". Suppose instead that we are continually making new observations and wish to update our current estimate. One possibility would be to recalculate our estimate at every step. If we are estimating a parameter using the sample mean, a little thought shows, the batch approach is not necessary:

$$\begin{aligned}
\bar{x}_{n+1} &= \frac{1}{n+1}\sum_{i=1}^{n+1} z_i \\
&= \frac{1}{n+1}\,(z_{n+1} + \sum_{i=1}^{n} z_i) \\
&= \bar{x}_n + \frac{1}{n+1}\,(z_{n+1} - \bar{x}_n)
\end{aligned}$$

Consider again the regression problem where our (scalar) observations are of the form

$$z_i = \mathbf{h}_i^\top \boldsymbol{\theta} + w_i$$

where $\mathbf{h}_i$ is a column vector, and w_i is zero mean gaussian noise with variance σ_i^2, so

$$z_i|\boldsymbol{\theta} \sim N(\mathbf{h}_i^\top \boldsymbol{\theta}, \sigma_i^2)$$

For example for cubic regression $\mathbf{h}_i^\top = \begin{bmatrix} 1 & x_i & x_i^2 & x_i^3 \end{bmatrix}$.

We can now consider formulating an *iterative* process where at iteration n we use the previous step's posterior density as the prior for the current one:

- prior:

$$\boldsymbol{\theta} \sim N(\hat{\boldsymbol{\theta}}_n, \mathbf{P}_n)$$

- likelihood:

$$z_n|\boldsymbol{\theta} \sim N(\mathbf{h}_n^\top \boldsymbol{\theta}, \sigma_n^2)$$

- posterior:

$$\begin{aligned}
\mathbf{P}_{n+1} &= (\mathbf{P}_n^{-1} + \mathbf{h}_n^\top \mathbf{h}_n/\sigma_n^2)^{-1} \\
\boldsymbol{\theta}_{n+1} &= \mathbf{P}_{n+1}\left[\mathbf{P}_n^{-1}\boldsymbol{\theta}_n + \mathbf{h}_n^\top z_n/\sigma_n^2\right]
\end{aligned}$$

We have a slight problem with this formulation since to begin with our covariance matrix will be "infinite" – i.e. zero information. Instead we can perform the iteration on the information matrix, the inverse of the covariance matrix $\mathbf{S} = \mathbf{P}^{-1}$, and on the **information weighted mean $\mathbf{S}\hat{\boldsymbol{\theta}}$**:

$$\begin{aligned}
\mathbf{S}_{n+1} &= \mathbf{S}_n + \mathbf{h}_n^\top \mathbf{h}_n/\sigma_n^2 \\
\mathbf{S}_{n+1}\boldsymbol{\theta}_{n+1} &= \mathbf{S}_n\boldsymbol{\theta}_n + \mathbf{h}_n^\top z_n/\sigma_n^2
\end{aligned}$$

With this iteration we can initialize the process with $\mathbf{S}_0 = 0, \mathbf{S}_0\hat{\boldsymbol{\theta}} = 0$, or with any prior model we may have, and obtain the solution by inverting $\mathbf{S}_n$ (which will hope by then will be non-singular). Thus (for independent gaussian errors) we obtain a MLE estimate of the regression parameters and a covariance on the regression parameters. If we have a prior on the regression parameters then the estimate is also MAP.

3.11 Stochastic difference equations

To this point we have considered estimating a set of unknown, fixed parameters. However, given that we can do this recursively, the obvious next step is to consider estimating a state which evolves over time.

Hence, we now return briefly to one of the motivating examples, that of feedback control. Recall that we modelled the system with state equations of the form (also recall figure 9):

$$\begin{aligned}\mathbf{x}(\mathbf{k+1}) &= \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k) + \mathbf{v}(k)\\ \mathbf{y}(k) &= \mathbf{C}\mathbf{x}(k) + \mathbf{w}(k)\end{aligned}$$

where the $\mathbf{v}(k)$ and $\mathbf{w}(k)$ are vectors of zero mean, independent random variables modelling stochastic exogenous variables and unknown constants ($\mathbf{v}$) and sensor noise ($\mathbf{w}$). The effect of the introduction of these stochastic variables means that the time dependent system variables such as the state $\mathbf{x}$ and the output $\mathbf{y}$ also become stochastic variables. These equations do not have unique solutions for (say) $\mathbf{x}(k)$ for all values of k as would the deterministic versions without $\mathbf{v}, \mathbf{w}$. Future values of the solution $\mathbf{x}(k)$ are random variables. Consider, for example a scalar difference equation:

$$x(k+1) = ax(k) + v(k), \quad p(v(k)) = N(0, \sigma^2)$$

Hence

- $p(x(k+1)|x(k)) = N(ax(k), \sigma^2)$
- $p(x(k+2)|x(k)) = N(a^2x(k), \sigma^2(1+a^2))$
- etc

The effect is that the future values become less and less certain. You can see this in figure 28 where a set of random walk paths has been generated using matlab:

```
function path(s0,v,s,N)  % s0: initial sd. v: vertical drift velocity
                         % s: random step sd.  N: no of steps
  x = [1:N]; y = [1:N];
  x(1)=s0*randn(1);      % initialise
  y(1)=s0*randn(1);
  for n=2:N
    x(n) = x(n-1) + s * randn(1);
    y(n) = y(n-1) + s * randn(1) + v;
  end;
  plot(x,y,x(1),y(1),'+',x(N),y(N),'*');


>> for i=1:100
>>    path(0,0.01,0.01,50);
>> end
```

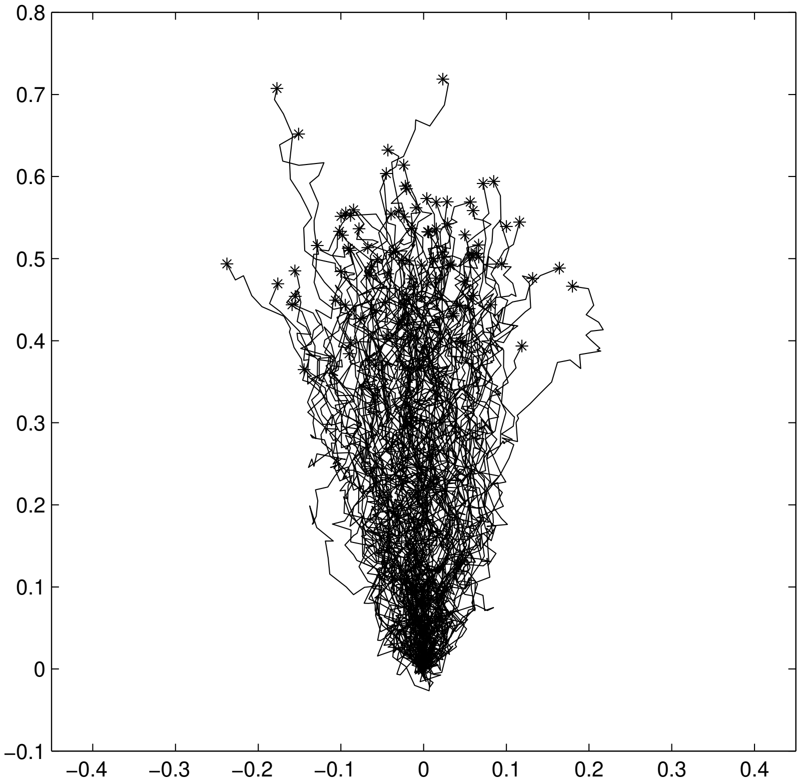


Figure 28: Random evolution of a simple stochastic difference equation

3.12 Kalman filter

In the second half of this course you will encounter the Kalman Filter in much more detail (and arrive at it via a slightly different path). However since we now have all the mathematical ideas in place, we might as well derive it from Bayes' rule.

Consider the general system above. At each time step the system state evolves according to:

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k) + \mathbf{v}$$

The idea is to make a **prediction** of the state according to the stochastic difference equations above. We can determine the p.d.f. of the prediction as

$$\begin{aligned} p(x(k+1)|x(k)) &= N(\mathbf{A}\hat{\mathbf{x}}(k|k) + \mathbf{B}\mathbf{u}(k), \mathbf{A}\mathbf{P}(k|k)\mathbf{A}^\top) \\ &= N(\hat{\mathbf{x}}(k+1|k), \mathbf{P}(k+1|k)) \end{aligned}$$

This becomes our **prior** for the update step.

Now we make a measurement and hence determine the **likelihood function** $p(z(k+1)|x(k+1))$:

$$p(z(k+1)|x(k+1)) = N(\mathbf{H}\hat{\mathbf{x}}(k+1|k), \mathbf{R})$$

We can now compute a posterior density by combining the prior and likelihood. Our previous results tell us that this will be normal, and will have mean and covariance:

$$\begin{aligned}
\mathbf{P}(k+1|k+1) &= (\mathbf{P}(k+1|k)^{-1} + \mathbf{H}^\top \mathbf{R}^{-1}\mathbf{H})^{-1} \\
\hat{\mathbf{x}}(k+1|k+1) &= \mathbf{P}(k+1|k+1)\left[\mathbf{P}(k+1|k)^{-1}\hat{\mathbf{x}}(k+1|k) + \mathbf{H}^\top \mathbf{R}^{-1}\mathbf{z}(k+1)\right]
\end{aligned}$$

which are the Kalman Filter update equations. The latter is more often seen in the form

$$\hat{\mathbf{x}}(k+1|k+1) = \hat{\mathbf{x}}(k+1|k) + \mathbf{P}(k+1|k+1)\mathbf{H}\mathbf{R}^{-1}\left[\mathbf{z}(k+1) - \mathbf{H}\hat{\mathbf{x}}(k+1|k)\right]$$

3.13 Further reading topics

- Principal components analysis
- Outlier rejection and robust estimation