Chapter 2

A brief review of probability and statistics

In the last chapter, we have seen that data assimilation involves the combination of information from various sources according to their accuracies. In essence, we are viewing the data sources being random, with a particular observation being a realization of a random variable. Before we can develop theories on how best to combine the data sources, we must first step back and review some basic facts about probability and random variables.

2.1 Probability

Although Todling (1999) provides a very mathematical definition of probability, it is sufficient to have an intuitive idea of the concept. The probability of an event, ω, is given by a number between 0 and 1. Let Ω be the sample space or the set of all possible events, $\Omega = \{\omega\}$.

Definition: A probability model consists of a specification of

1. the space Ω of all possible outcomes, one of which is ω
2. a class of subsets of Ω called events, such that if A and B are events, then so are

$$A \cup B, \qquad A \cap B, \qquad A'$$

   where $A' = \Omega - A$ is the complement of A.

3. an assignment of a number $P(A)$ to each event A satisfies

$$0 \leq P(A) \leq 1$$
$$P(\Omega) = 1$$
$$P(A \cup B) = P(A) + P(B)$$

   if A and B are disjoint events.

Now that we have defined the probability function, P, it will be extremely useful to define a *conditional* probability. Suppose an event B has occurred. The conditional probability is concerned with the probability of A given that B has occurred.

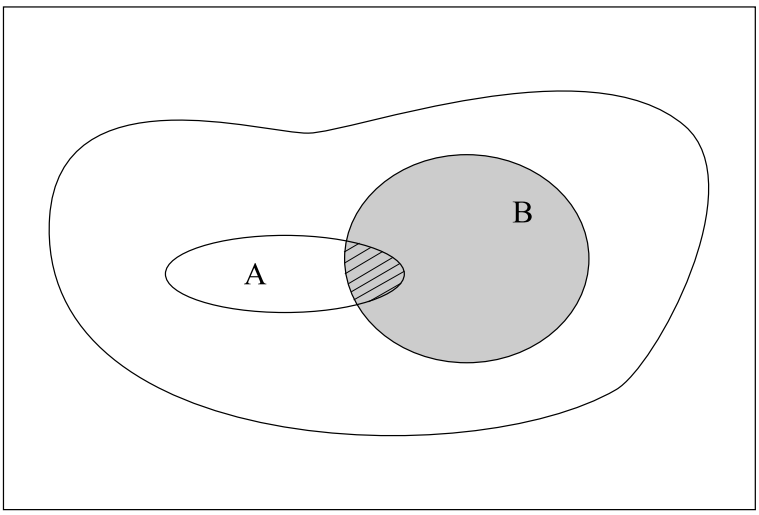

Figure 2.1: Conditional probability. $P(A|B) = P(A \cap B)/P(B)$. If event B has already occurred, we must be in the shaded region. Then the probability of A occurring is the probability of the striped region over the probability of the shaded region.

Definition: The probability of the event A given that B has occurred is denoted by $P(B|A)$ and defined by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Intuitively, this makes sense. Consider Fig. 2.1. If event B has occurred, we are in the shaded region. Then $P(A|B)$ must be the probability of the striped region over the probability of the shaded region.

Example 2.1 *Urn*

Suppose we have an urn with 4 balls, 2 red and 2 black, and we draw 2 balls without replacing them. (1) What is the joint probability of the 1st and 2nd draws? (2) What is the conditional probability of the 2nd ball being red given that the first one is red?

(1) There are 4 possible outcomes to the experiment. The joint probability of the two draws is as follows:

$p(Red \quad Red) = \frac{1}{2}\frac{1}{3} = \frac{1}{6}$
$p(Blue \quad Blue) = \frac{1}{2}\frac{1}{3} = \frac{1}{6}$
$p(Red \quad Blue) = \frac{1}{2}\frac{2}{3} = \frac{2}{6}$
$p(Blue \quad Red) = \frac{1}{2}\frac{2}{3} = \frac{2}{6}$

(2) If the first draw is red, 2 blues and 1 red remain so P(red)=1/3. Another way to do this is to let A be the event of a red on the first draw and B be the event of a red on the second draw. Then $P(B|A) = P(A \cap B)/P(A) = (1/6)/(1/2) = 1/3$.

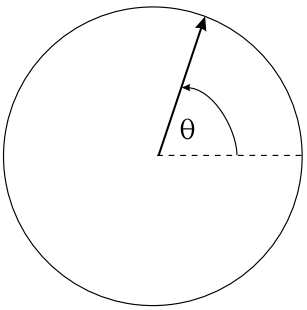

Figure 2.2: Example of a random variable. In a game, a pointer is used. The location of the point can be mapped to the real line by using the angle of the pointer.

2.2 Random variables

Measurements contain noise so that if an observation were repeated, different values could result. We can use the concept of a random variable to represent this effect. A real *random variable* or r.v. is simply a function which maps the set of events to the real line. More precisely, a real random variable, x, is a real function whose domain is the space Ω such that a) the set $\{\mathrm{x} \leq x\}$ is an event for any real number, x, and b) the probability of the events $P(\mathrm{x} = +\infty) = P(\mathrm{x} = -\infty) = 0$. Note that x is a possible value of the random variable x called a realization. We shall use the math-italic font to indicate the realization, and the Roman font to indicate the variable.

Example 2.2 *Card Game (Brown p.17)*

Suppose we have a deck of cards and a game applies the following values to each card or outcome:

Card	*Value*
2-9	*0*
10	*10*
Jack	*1*
Queen	*2*
King	*3*
Ace	*4*

Here we are mapping the sample space of $\{\mathrm{Ace}, 2, 3, \ldots, 10, \mathrm{Jack}, \mathrm{Queen}, \mathrm{King}\}$ *to the numbers* $\{0, 1, 2, 3, 4, 10\}$. *This mapping is not one-to-one.*

Example 2.3 *Pointer (Brown p. 18)*

The outcome of an experiment is the location of a pointer (see Fig. 2.2). We can define the location by the angle of the pointer. Then the sample space is the infinity of points on the circle. The values that the random variable can take are $[0, 2\pi)$.

2.2.1 Probability density functions

A random variable is completely described by its **probability distribution function**. For a continuous r.v., the probability distribution function, $F_{\mathrm{x}}(x)$, is given by:

$$F_{\mathrm{x}}(x) = P(\mathrm{x} \leq x) = \int_{-\infty}^{x} p_{\mathrm{x}}(s)ds. \tag{2.1}$$

(It is only continuous r.v.s that we will encounter in this course.) The function $p_{\mathrm{x}}(x)$ is the **probability density function** or p.d.f. From this definition and the fact that $0 \leq p_{\mathrm{x}}(x) \leq 1$, it is clear that $F_{\mathrm{x}}(x)$ is a non-decreasing function and that $F_{\mathrm{x}}(-\infty) = 0$ and $F_{\mathrm{x}}(+\infty) = 1$. In addition,

$$\int_{-\infty}^{+\infty} p_{\mathrm{x}}(s)ds = 1.$$

It turns out that we almost always deal with p.d.f.'s rather than with probability distribution functions. In fact, often when we refer to an underlying distribution of a process, we are actually speaking of the the p.d.f. of the r.v., since one may often be deduced from the other.

Example 2.4 *Card game.*

In the above card game example, let us assume equal probabilities for all elements in the sample space. Then,

ω	$x(\omega)$	$p(x)$
2-9	*0*	*32/52*
10	*10*	*4/52*
Jack	*1*	*4/52*
Queen	*2*	*4/52*
King	*3*	*4/52*
Ace	*4*	*4/52*

This is an example of a discrete r.v. where all of the possible realizations are countable. As noted above, in this course, we will deal almost exclusively with continuous r.v.s.

Example 2.5 *Pointer.*

In the above pointer example, the probability of landing on any particular point is infinitessimal. If all points are equally likely, then

$$F_{\mathrm{x}}(\theta) = P(\mathrm{x} \leq \theta) = \begin{cases} 0 & \theta < 0 \\ \theta/(2\pi) & 0 \leq \theta < 2\pi \\ 1 & \theta > 2\pi \end{cases} \tag{2.2}$$

is the probability distribution function. Clearly, $F_{\mathrm{x}}(\theta) \to 0$ as $\theta \to -\infty$ and $F_{\mathrm{x}}(\theta) \to 1$ as $\theta \to +\infty$ and $F_{\mathrm{x}}(\theta)$ is a nondecreasing function of θ. The p.d.f. is given by

$$p_{\mathrm{x}}(\theta) = \begin{cases} 0 & \theta < 0 \\ 1/(2\pi) & 0 \leq \theta \leq 2\pi \\ 1 & \theta > 2\pi \end{cases} \tag{2.3}$$

This is an example of a uniform distribution function.

2.2.2 Normal distribution

An r.v. has a normal or Gaussian distribution if its p.d.f. is given by

$$p_{\mathrm{x}}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{2.4}$$

where

$$\mu = \int_{-\infty}^{+\infty} x p_{\mathrm{x}}(x) dx$$

$$\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 p_{\mathrm{x}}(x) dx.$$

The normal density is important because it has some very nice mathematical properties. One important property is that the normal density is completely defined by its mean and variance and is written

$$\mathrm{x} \sim \mathcal{N}(\mu, \sigma^2).$$

Variance is a measure of dispersion about the peak so that a small σ^2 corresponds to a sharp peak while a large σ^2 corresponds to a flat peak. The normal distribution is also important because it appears quite often in practice. One possible reason for this is given by the "Central Limit Theorem" which says that a super-position of independent random variables always tends toward normality regardless of the individual distributions involved. Therefore, if noise is due to a superposition of many small contributions, it may be reasonable to assume normality.

2.2.3 Moments of a distribution

It is very difficult to determine the complete p.d.f. of a random variable, in practice. Often, it may be sufficient to determine the properties of the p.d.f. Other times, these properties may simply be the only information which can be practically obtained.

The **mean** or expected value of x is

$$\mu = \mathcal{E}\{x\} = \int_{-\infty}^{+\infty} x p(x) dx.$$

In general, for any function, $f(x)$:

$$\mathcal{E}\{f(x)\} = \int_{-\infty}^{+\infty} f(x) p(x) dx.$$

If $f(x) = x^n$, then

$$\mathcal{E}\{x^n\} = \int_{-\infty}^{+\infty} x^n p(x) dx$$

defines the nth moment of **x**. The first moment is the mean. The nth central moment is defined by

$$\mathcal{E}\{(x - \mathcal{E}(x))^n\} = \int_{-\infty}^{+\infty} (x - \mathcal{E}(x))^n p(x) dx.$$

The second central moment is the **variance**:

$$\mathrm{var}(x) = \mathcal{E}\{(x - \mathcal{E}(x))^2\} = \mathcal{E}\{x^2\} - (\mathcal{E}\{x\})^2.$$

The **standard deviation** is simply the square root of the variance:

$$\sigma = \sqrt{\text{var}(x)}.$$

Very often it is not convenient or even possible to compute an entire p.d.f. of a random variable, so we are content with some information about the p.d.f. such as the mean. There are, however, other possibilities such as the **median**, the **mid-range** and the **mode**. The median, μ_1, divides the area under a p.d.f. in half:

$$\mu_1 = \int_{-\infty}^{\mu_1} p(x)dx = \int_{\mu_1}^{\infty} p(x)dx = \frac{1}{2}.$$

The mid-range is simply:

$$\mu_p = 0.5 * [\max_{\mathrm{x}}(\mathrm{x}) + \min_{\mathrm{x}}(\mathrm{x})].$$

The mode is the most likely value. For a p.d.f. with many peaks, there may be many modes. Such a distribution is called multi-modal. The mode is defined by

$$\left.\frac{dp_{\mathrm{x}}(x)}{d\mathrm{x}}\right|_{x=m} = 0.$$

For a Normal p.d.f., the mean, median and mode are identical. For a symmetric p.d.f., the mean equals the median. The proof of this is given in problem 2.1. For a symmetric and unimodal p.d.f., the mean, median and mode are identical.

Example 2.6 *Pointer.*

The p.d.f. for this example is

$$p(x) = \begin{cases} \frac{1}{2\pi} & 0 \le x < 2\pi \\ 0 & \text{otherwise} \end{cases}.$$

The mean is

$$\mathcal{E}\{x\} = \int_0^{2\pi} x\frac{1}{2\pi}dx = \frac{1}{2\pi}\,\frac{1}{2}x^2\Big|_0^{2\pi} = \frac{1}{4\pi}(4\pi^2 - 0) = \pi. \tag{2.5}$$

The variance is

$$\begin{aligned} \text{var}(x) &= \mathcal{E}\{x^2\} - (\mathcal{E}\{x\})^2 = \int_0^{2\pi} x^2\frac{1}{2\pi}dx - \pi^2 \\ &= \frac{1}{2\pi}\,\frac{1}{3}x^3\Big|_0^{2\pi} - \pi^2 = \frac{1}{6\pi}(8\pi^3 - 0) - \pi^2 = \frac{1}{3}\pi^2. \end{aligned} \tag{2.6}$$

The standard deviation is $\frac{\pi}{\sqrt{3}}$.

2.2.4 Characteristic Functions

Another way to represent an r.v. is by its characteristic function:

$$\phi_{\mathrm{x}}(u) = \mathcal{E}\{e^{iu\mathrm{x}}\} = \int_{-\infty}^{\infty} e^{iux}p_{\mathrm{x}}(x)dx. \tag{2.7}$$

$\phi_{\mathrm{x}}(u)$ is just the Fourier transform of the p.d.f. The characteristic function is especially useful in evaluating the moments of x:

$$\mathcal{E}\{\mathrm{x}^n\} = \frac{1}{i}\,\left.\frac{d^n\phi_{\mathrm{x}}(u)}{du^n}\right|_{u=0}. \tag{2.8}$$

2.3 Multivariate random variables

Let us reconsider the simple scalar example of Chapter 1. What if we wanted to simultaneously find the analysis of temperature at Toronto and Montreal? We could define our background information as $\mathbf{x}^b = (\mathrm{x}_T^b, \mathrm{x}_M^b)^{\mathrm{T}}$, where the subscripts T and M refer to Toronto and Montreal. Similarly, we'd need an observation vector, $\mathbf{x}^{\mathrm{obs}} = (\mathrm{x}_T^{\mathrm{obs}}, \mathrm{x}_M^{\mathrm{obs}})^{\mathrm{T}}$ and an analysis vector, $\mathbf{x}^a = (\mathrm{x}_T^a, \mathrm{x}_M^a)^{\mathrm{T}}$. As in Chapter 1, we can choose to combine our two sources of information linearly, ie.

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{W}(\mathbf{x}^{\mathrm{obs}} - \mathbf{x}^b). \tag{2.9}$$

Note the similarity between (1.3) and (2.9). We have simply replaced scalar quantities with vectors, and the weight is now a 2x2 matrix. Since we will interpret these quantities as random variables, it is clear that we will primarily be dealing with multivariate random variables. In this simple example, $\mathbf{x}$ refers to a single variable, such as temperature, but at different spatial locations. More generally, $\mathbf{x}$ can refer to all prognostic forecast variables such as temperature, zonal and meridional wind components, surface pressure, specific humidity. In this case, $\mathbf{x}$ is a vector of length equal to the number of variables times the number of gridpoints. In this section, we will generalize the concepts of probability density functions for vector random variables.

2.3.1 Joint density and distribution functions

The r.v.s x and y are jointly distributed if they are defined in the same probability space. The joint p.d.f. is then $p_{\mathrm{xy}}(x, y)$. The marginal densities of x and y are

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

$$p_{\mathrm{y}}(y) = \int_{-\infty}^{\infty} p_{\mathrm{xy}}(x, y)dx.$$

Example 2.7 *Coins (Brown p30)*

A sack contains 2 pennies, 1 nickel and 1 dime. Draw 1 coin and replace it. Let x *be the value of the first coin drawn. Let* y *be the value of the second coin drawn. What is the discrete joint p.d.f. of* x *and* y*?*

	1	*5*	*10*	$p(x_i)$
1	$\frac{1}{2}\frac{1}{2}$	$\frac{1}{2}\frac{1}{4}$	$\frac{1}{2}\frac{1}{4}$	$\frac{1}{2}$
5	$\frac{1}{4}\frac{1}{2}$	$\frac{1}{4}\frac{1}{4}$	$\frac{1}{4}\frac{1}{4}$	$\frac{1}{4}$
10	$\frac{1}{4}\frac{1}{2}$	$\frac{1}{4}\frac{1}{4}$	$\frac{1}{4}\frac{1}{4}$	$\frac{1}{4}$
$p(y_i)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	

The Marginal p.d.f.s are

$$p_{\mathrm{x}}(x_i) = \sum_j p_{\mathrm{xy}}(x_i, y_j)$$

$$p_{\mathrm{y}}(y_j) = \sum_i p_{\mathrm{xy}}(x_i, y_j)$$

The marginal p.d.f. $p_{\mathrm{x}}(x_i)$ *represents the probability of getting a specific coin first. The probability of getting a penny first is 1/2, from the above table, and the probability of getting a nickel or a dime first is 1/4. Similarly,* $p_{\mathrm{y}}(y_j)$ *is the probability of getting a certain coin on the second draw. Note that in this example,* $p_{\mathrm{xy}}(x_i, y_j) = p_{\mathrm{x}}(x_i)p_{\mathrm{y}}(y_j)$.

x, y are statistically independent if $p_{\rm xy}(x_i, y_j) = p_{\rm x}(x_i)p_{\rm y}(y_j)$ for all $\rm x_i$, $\rm y_j$.

Example 2.8 *Darts (Brown, p35)*

Let the position of a hit be given by its (x,y) coordinates. After sufficient practice, assume that the scatter in position is unbiased in both directions. The joint p.d.f. is then

$$p_{\rm xy}(x, y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}. \tag{2.10}$$

The marginal densities are:

$$\begin{aligned} p_{\rm x}(x) &= \int_{-\infty}^{\infty} p_{\rm xy}(x, y)dy = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} \\ p_{\rm y}(y) &= \int_{-\infty}^{\infty} p_{\rm xy}(x, y)dx = \frac{1}{\sqrt{2\pi}\sigma} e^{-y^2/2\sigma^2}. \end{aligned}$$

The expectation of the product of two r.v.s is of special interest:

$$\mathcal{E}({\rm xy}) = \int_{-\infty}^{\infty} xy p_{\rm xy}(x, y)dxdy. \tag{2.11}$$

If x and y are independent, then

$$\mathcal{E}({\rm xy}) = \int_{-\infty}^{\infty} x p_{\rm x}(x)dx \int_{-\infty}^{\infty} y p_{\rm y}(y)dy = \mathcal{E}(x)\mathcal{E}(y). \tag{2.12}$$

If $\mathcal{E}({\rm xy}) = \mathcal{E}({\rm x})\mathcal{E}({\rm y})$, then x, y are uncorrelated.

If x, y are independent, then they are uncorrelated.

If x, y are uncorrelated, **this does not mean that** they are independent.

If $\mathcal{E}({\rm xy}) = 0$, x and y are said to be orthogonal.

The **covariance** of x and y is

$$cov({\rm x}, {\rm y}) = \mathcal{E}\{({\rm x} - \mu_x)({\rm y} - \mu_y)\}.$$

The correlation coefficient is defined as

$$\rho = \frac{\mathcal{E}\{({\rm x} - \mu_x)({\rm y} - \mu_y)\}}{\sqrt{\mathcal{E}\{({\rm x} - \mu_x)^2\}}\sqrt{\mathcal{E}\{({\rm y} - \mu_y)^2\}}}.$$

Note that if x=y, then ρ=1, and if x=-y, ρ=-1. If x and y are uncorrelated, then

$$\begin{aligned} \mathcal{E}\{({\rm x} - \mu_x)({\rm y} - \mu_y)\} &= \mathcal{E}\{{\rm xy} - \mu_x{\rm y} - \mu_y{\rm x} + \mu_x\mu_y\} \\ &= \mathcal{E}\{{\rm xy}\} - \mu_x\mu_y - \mu_y\mu_x + \mu_x\mu_y \\ &= \mathcal{E}\{{\rm x}\}\mathcal{E}\{{\rm y}\} - \mu_x\mu_y \\ &= \mu_x\mu_y - \mu_x\mu_y \\ &= 0. \end{aligned} \tag{2.13}$$

Thus if x and y are uncorrelated, then ρ=0.

2.3.2 Transformation of random variables

In data assimilation we are primarily dealing with 2 random variables, the model state, $\mathbf{x}$, and the observations, $\mathbf{y}$. In order to combine this information, we are going to need to compare these quantities. The relation between the two is defined by the **measurement equation**:

$$\mathbf{y} = \mathbf{H}(\mathbf{x}).$$

$\mathbf{H}$ is the observation operator and it maps the model state to the observed variables. For example, the model state variables might include wind and temperature, but the observed variables might be radiance or backscatter. Now with this relationship, and the knowledge of the p.d.f. of one of these variables, how can one deduce the p.d.f. of the other one? In this section, we answer this question for the case of two jointly distributed scalar variables.

If y = f(x) and we know the p.d.f. of x, what is the p.d.f. of y? Consider Fig. 2.3. We know that

$$P(x_0 \leq x \leq x_0 + \Delta x) = P(y_0 \leq y \leq y_0 + \Delta y).$$

Thus relating this to densities gives

$$\int_{x_0}^{x_0+\Delta x} p_x(x)dx = \begin{cases} \int_{y_0}^{y_0+\Delta y} p_y(y)dy & \text{if } \Delta y > 0 \\ \int_{y_0+\Delta y}^{y_0} p_y(y)dy = -\int_{y_0}^{y_0+\Delta y} p_y(y)dy & \text{if } \Delta y < 0 \end{cases} \tag{2.14}$$

Combining this information, and equating the integrands, we can write

$$p_x(x)dx = p_y(y)|dy| \tag{2.15}$$

or equivalently,

$$p_y(y) = p_x(x)\left|\frac{df^{-1}(y)}{dy}\right|. \tag{2.16}$$

Thus to convert the p.d.f. of x to a p.d.f. of y, simply write the p.d.f. of x in terms of y and multiply by the jacobian of the transformation.

Example 2.9 *If* x *is* $\mathcal{N}(0, \sigma^2)$ *find the p.d.f. of* y *where* $y = x^3$. *Since* $x = y^{1/3}$ *we have that*

$$|dx/dy| = |\frac{1}{3}y^{-2/3}| = \frac{1}{3y^{2/3}}.$$

Thus we can immediately write that

$$\begin{aligned} p_y(y) &= \frac{1}{3y^{2/3}}\frac{1}{\sigma\sqrt{2\pi}}e^{-x^2/(2\sigma^2)} \\ &= \frac{1}{3y^{2/3}}\frac{1}{\sigma\sqrt{2\pi}}e^{-y^{2/3}/(2\sigma^2)}. \end{aligned}$$

Thus although x is normally distributed, y is not. A nonlinear transformation of a Gaussian is not necessarily a Guassian.

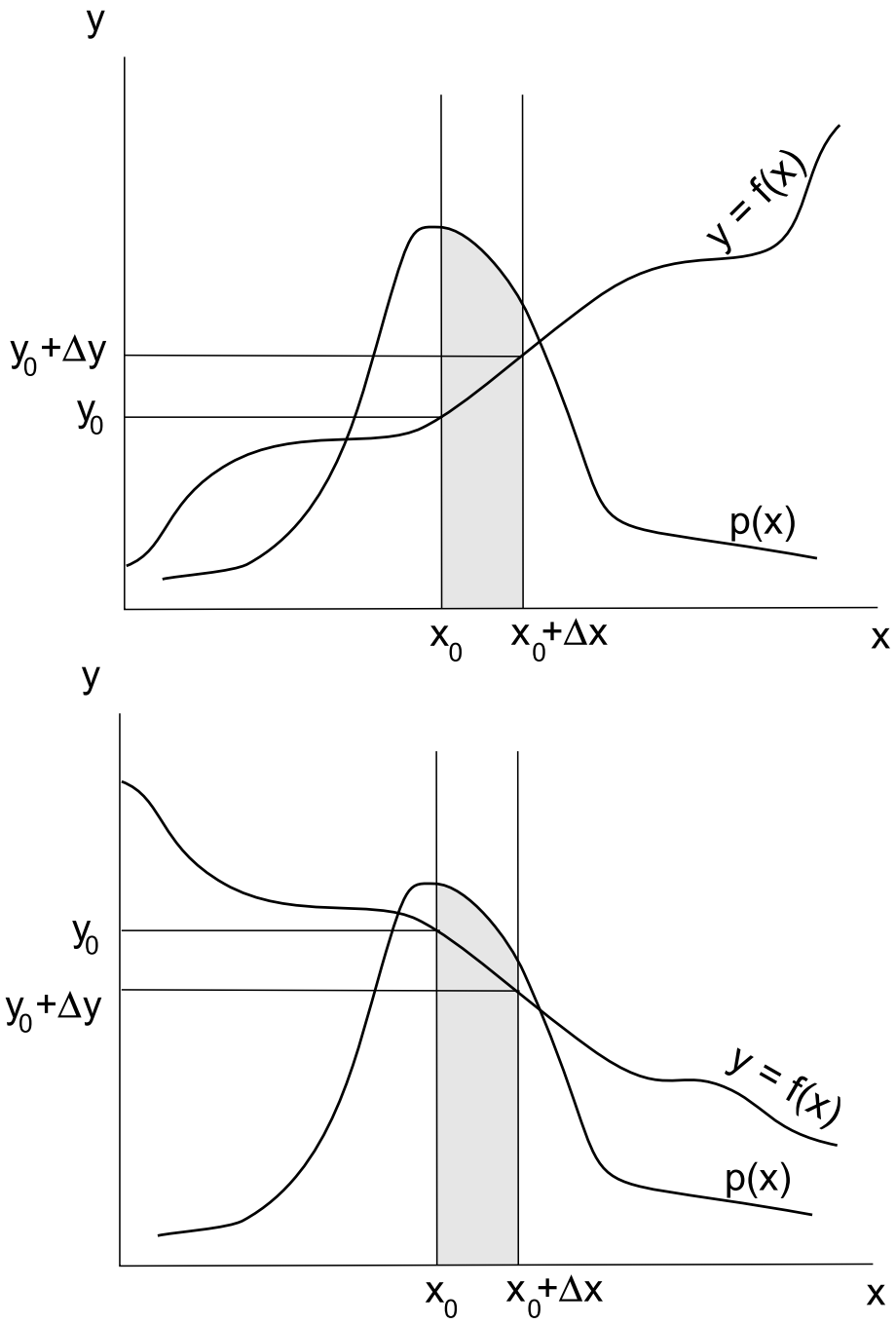

Figure 2.3: Transformation of a random variable. Given the p.d.f. of x and the relation y=f(x), what is the p.d.f. of y? For a small variation Δx about a point, x_0, the corresponding variation of y can be deduced using y=f(x). The top panel corresponds to the case where $\Delta y > 0$ while the bottom panel is for the case that $-\Delta y < 0$.

Example 2.10 *Darts*

Recall from example 2.8 involving a dart game, the probability of the hit location was

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

This is a special case of a bivariate Normal density where the two variates have 0 means and equal variances. What is the p.d.f. in polar coordinates?

We have that

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

The Jacobian of the transformation is

$$J\left(\frac{x,y}{r,\theta}\right) = \left| \begin{array}{cc} \frac{\partial x}{\partial r} & \frac{\partial y}{\partial r} \\ \frac{\partial x}{\partial \theta} & \frac{\partial y}{\partial \theta} \end{array} \right| = \left| \begin{array}{cc} \cos\theta & \sin\theta \\ -r\sin\theta & r\cos\theta \end{array} \right| = r. \tag{2.17}$$

Now, for an closed contour C in the x-y plane we have that

$$\int\int_C p_{\mathrm{xy}}(x,y)dxdy = \int\int_{C'} p_{\mathrm{xy}}(x(r,\theta),y(r,\theta))J\left(\frac{x,y}{r,\theta}\right)drd\theta = \int\int_{C'} p_{\mathrm{r}\theta}(r,\theta)drd\theta. \tag{2.18}$$

On equating the integrands of the last equality, we deduce that

$$p_{\mathrm{xy}}(x(r,\theta),y(r,\theta))J\left(\frac{x,y}{r,\theta}\right) = p_{\mathrm{r}\theta}(r,\theta). \tag{2.19}$$

Substituting for our given p.d.f. gives

$$\begin{aligned} p_{\mathrm{r}\theta}(r,\theta) &= \frac{r}{2\pi\sigma^2} e^{-((r\cos\theta)^2+(r\sin\theta)^2)/(2\sigma^2)} \\ &= \frac{r}{2\pi\sigma^2} e^{-r^2/(2\sigma^2)}. \end{aligned} \tag{2.20}$$

The marginal distributions are

$$p_{\mathrm{r}}(r) = \int_0^{2\pi} \frac{r}{2\pi\sigma^2} e^{-r^2/(2\sigma^2)} d\theta = \frac{r}{\sigma^2} e^{-r^2/(2\sigma^2)}, \tag{2.21}$$

$$p_\theta(\theta) = \int_0^{\infty} \frac{r}{2\pi\sigma^2} e^{-r^2/(2\sigma^2)} dr = \frac{1}{2\pi}(-e^{-r^2/(2\sigma^2)})\Big|_0^{\infty} = \frac{1}{2\pi}. \tag{2.22}$$

Thus similar, independent zero mean Normal densities in x, y correspond to Rayleigh and uniform densities in the r, θ domain. How do we interpret this, in terms of the dart game? Because the player is very good and there was no bias in the hit location in either x or y, the probability of hitting a particular angle is the same as that for any other angle between 0 and 2π. The probability of hitting the bullseye is identically zero. The probability of hitting within a radius r increases as r increases until r becomes infinite. Of course, the probability of hitting within an infinite distance of the bullseye is 1. When r< σ, the probability increases almost linearly with distance. When r=σ the rate of increase in probability diminishes.

2.3.3 Vector Notation

Let us define an n-vector, $\mathbf{x}$ as

$$\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n)^{\mathrm{T}}.$$

A realization of the random vector, $\mathbf{x}$ is then

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}.$$

As noted in Todling (1999), all of the previous definitions can be written in vector notation. For example, the p.d.f. is

$$p_{\mathbf{x}}(\boldsymbol{x}) = p_{\mathrm{x}_1\mathrm{x}_2\ldots\mathrm{x}_n}(x_1, x_2, \ldots, x_n).$$

The probability distribution is

$$\begin{aligned} F_{\mathbf{x}}(\boldsymbol{x}) &= \int_{-\infty}^{\boldsymbol{x}} p_{\mathbf{x}}(\boldsymbol{x}')d\boldsymbol{x}' \\ &= \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_n} p_{\mathrm{x}_1\ldots\mathrm{x}_n}(x_1', \ldots, x_n')dx_1' \ldots dx_n'. \end{aligned} \tag{2.23}$$

The probability density function is defined as the derivative of the distribution function:

$$p_{\mathbf{x}}(\boldsymbol{x}) = \frac{\partial^n F_{\mathbf{x}}(\boldsymbol{x})}{\partial \boldsymbol{x}} = \frac{\partial^n F_{\mathbf{x}}(\boldsymbol{x})}{\partial x_1 \ldots \partial x_n}. \tag{2.24}$$

The expected value of the random vectors is defined as

$$\mathcal{E}(\mathbf{x}) = (\mathcal{E}(\mathrm{x}_1), \ldots, \mathcal{E}(\mathrm{x}_n))^{\mathrm{T}}.$$

The **covariance matrix** is formed by the outer product of the vector $\mathbf{x} - \mathcal{E}(\mathbf{x})$ with itself:

$$\mathbf{P}_{\mathbf{x}} = \mathcal{E}\{(\mathbf{x} - \mathcal{E}(\mathbf{x}))(\mathbf{x} - \mathcal{E}(\mathbf{x}))^{\mathrm{T}}\} \tag{2.25}$$

or, in more detail,

$$\mathbf{P}_{\mathbf{x}} = \begin{bmatrix} var(\mathrm{x}_1) & cov(\mathrm{x}_1, \mathrm{x}_2) & \ldots & cov(\mathrm{x}_1, \mathrm{x}_n) \\ cov(\mathrm{x}_2, \mathrm{x}_1) & var(\mathrm{x}_2) & \ldots & cov(\mathrm{x}_2, \mathrm{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\mathrm{x}_n, \mathrm{x}_1) & cov(\mathrm{x}_n, \mathrm{x}_2) & \ldots & var(\mathrm{x}_n) \end{bmatrix}. \tag{2.26}$$

Although there are any number of higher order moments, in practice, they are difficult to estimate so we usually deal only with the first two moments: the mean and covariance. If one were to consider a model state for a forecast model, then $\mathbf{x}$ could be of dimension 10^7 (assuming 5 variables, 200x400x28 gridpoints). The covariance matrix would then be 10^7 x 10^7! Computationally this is very difficult to handle. Moreover, given the number of observations typically available in 6 hours, we could not determine all of the matrix elements. Thus, in practice we try to simplify or model the forecast error covariance to reduce the number of parameters to estimate to a number small enough to be determined from the available observations.

Note that the covariance matrix is a symmetric, positive definite matrix. The positive definite part means that

$$\mathbf{y}^{\mathrm{T}}\mathbf{P}_{\mathbf{x}}\mathbf{y} > 0, \forall \mathbf{y} \in \mathcal{R}^n.$$

It is not hard to prove that a covariance matrix $\mathbf{P}$ is positive semi-definite. The following proof is from Daley (1991), pg. 418.

Definition: A matrix $\mathbf{P}$ $(n \times n)$ is positive definite if for all $\mathbf{y}$ $(n \times 1)$,

$$\mathbf{y}^{\mathrm{T}}\mathbf{P}\mathbf{y} > 0.$$

$\mathbf{P}$ is positive *semi*-definite if for all $\mathbf{y}$,

$$\mathbf{y}^{\mathrm{T}}\mathbf{P}\mathbf{y} \geq 0.$$

To complete the proof, expand the expression in terms of components of the vector, $\mathbf{y} = (y_1, y_2, \ldots, y_n)^{\mathrm{T}}$:

$$\begin{aligned}
\mathbf{y}^{\mathrm{T}}\mathbf{P}\mathbf{y} &= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{y}_i \mathbf{P}_{i,j} \mathrm{y}_j \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{y}_i E[(\mathrm{x}_i - E(\mathrm{x}_i))(\mathrm{x}_j - E(\mathrm{x}_j))]\mathrm{y}_j \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} E[\mathrm{y}_i(\mathrm{x}_i - E(\mathrm{x}_i))(\mathrm{x}_j - E(\mathrm{x}_j))\mathrm{y}_j] \\
&= E[\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{y}_i(\mathrm{x}_i - E(\mathrm{x}_i))(\mathrm{x}_j - E(\mathrm{x}_j))\mathrm{y}_j] \\
&= E[|\sum_{i=1}^{n} \mathrm{y}_i(\mathrm{x}_i - E(\mathrm{x}_i))|^2] \\
&\geq 0.
\end{aligned}$$

The interesting thing about real, symmetric positive semi-definite matrices is that they have some nice mathematical properties. Specifically, their eigenvalues are real and non-negative. This fact will be used later on.

Since Gaussians will be used often in this course, it is useful to note the multivariate Gaussian distribution. If $\mathbf{x}$, an n-vector, is $\mathcal{N}(\boldsymbol{\mu}, \mathbf{P})$, its p.d.f. can be written as

$$p_{\mathbf{x}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\mathbf{P}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{P}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right]. \tag{2.27}$$

In general, the joint characteristic function can be written as

$$\phi_{\mathbf{x}}(\mathbf{u}) = \mathcal{E}(\exp(i\mathbf{u}^{\mathrm{T}}\mathbf{x})). \tag{2.28}$$

Todling (1999) p. 10 shows that for a multivariate Gaussian, the characteristic function is

$$\phi_{\mathbf{x}}(\mathbf{u}) = \exp(i\mathbf{u}^{\mathrm{T}}\boldsymbol{\mu} - \frac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{P}^{-1}\mathbf{u}). \tag{2.29}$$

As noted above, it will be very useful to find the p.d.f. of a vector $\mathbf{y}$ given the p.d.f of $\mathbf{x}$ and the relationship between the two, $\mathbf{y} = \mathbf{f}(\mathbf{x})$:

$$p_{\mathbf{y}}(\boldsymbol{y}) = p_{\mathbf{x}}(\mathbf{f}^{-1}(\mathbf{y}))|\mathrm{Jac}(\mathbf{f}^{-1}(\mathbf{y})|. \tag{2.30}$$

2.3.4 2D Multivariate Gaussian

Because of the importance of Gaussian densities in this course, let us examine a special (simple) case of a multivariate Gaussian— when the vector has dimension 2. Thus, let $\boldsymbol{x} = (x, y)^{\mathrm{T}}$. First, recall the definition of the Gaussian p.d.f.:

$$p_{\mathbf{x}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\mathbf{P}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{P}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]. \tag{2.31}$$

Now let's expand to find the covariance matrix in detail.

$$\mathbf{P} = \mathcal{E}\left\{\begin{pmatrix} \mathrm{x}-\mu_{\mathrm{x}} \\ \mathrm{y}-\mu_{\mathrm{y}} \end{pmatrix}(\mathrm{x}-\mu_{\mathrm{x}}, \mathrm{y}-\mu_{\mathrm{y}})\right\} = \begin{bmatrix} var(\mathrm{x}) & cov(\mathrm{x},\mathrm{y}) \\ cov(\mathrm{x},\mathrm{y}) & var(\mathrm{y}) \end{bmatrix} = \begin{bmatrix} \sigma_{\mathrm{x}}^2 & \rho\sigma_{\mathrm{x}}\sigma_{\mathrm{y}} \\ \rho\sigma_{\mathrm{x}}\sigma_{\mathrm{y}} & \sigma_{\mathrm{y}}^2 \end{bmatrix}. \tag{2.32}$$

The determinant of $\mathbf{P}$ is

$$|\mathbf{P}| = \sigma_{\mathrm{x}}^2\sigma_{\mathrm{y}}^2 - \rho^2\sigma_{\mathrm{x}}^2\sigma_{\mathrm{y}}^2 = (1-\rho^2)\sigma_{\mathrm{x}}^2\sigma_{\mathrm{y}}^2$$

so that the inverse of $\mathbf{P}$ is

$$\mathbf{P}^{-1} = \frac{1}{(1-\rho^2)\sigma_{\mathrm{x}}^2\sigma_{\mathrm{y}}^2}\begin{bmatrix} \sigma_{\mathrm{y}}^2 & -\rho\sigma_{\mathrm{x}}\sigma_{\mathrm{y}} \\ -\rho\sigma_{\mathrm{x}}\sigma_{\mathrm{y}} & \sigma_{\mathrm{x}}^2 \end{bmatrix} = \frac{1}{1-\rho^2}\begin{bmatrix} \frac{1}{\sigma_{\mathrm{x}}^2} & -\frac{\rho}{\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}} \\ -\frac{\rho}{\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}} & \frac{1}{\sigma_{\mathrm{y}}^2} \end{bmatrix} \tag{2.33}$$

and the argument of the exponent in (2.31) (apart from a scalar) is

$$\begin{aligned}
(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{P}^{-1}(\mathbf{x}-\boldsymbol{\mu}) &= (\mathrm{x}-\mu_{\mathrm{x}}, \mathrm{y}-\mu_{\mathrm{y}})\frac{1}{1-\rho^2}\begin{bmatrix} \frac{1}{\sigma_{\mathrm{x}}^2} & -\frac{\rho}{\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}} \\ -\frac{\rho}{\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}} & \frac{1}{\sigma_{\mathrm{y}}^2} \end{bmatrix}\begin{pmatrix} \mathrm{x}-\mu_{\mathrm{x}} \\ \mathrm{y}-\mu_{\mathrm{y}} \end{pmatrix} \\
&= \frac{1}{1-\rho^2}\begin{bmatrix} \frac{\mathrm{x}-\mu_{\mathrm{x}}}{\sigma_{\mathrm{x}}^2} - \frac{\rho(\mathrm{y}-\mu_{\mathrm{y}})}{\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}} & -\frac{\rho(\mathrm{x}-\mu_{\mathrm{x}})}{\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}} + \frac{\mathrm{y}-\mu_{\mathrm{y}}}{\sigma_{\mathrm{y}}^2} \end{bmatrix}\begin{pmatrix} \mathrm{x}-\mu_{\mathrm{x}} \\ \mathrm{y}-\mu_{\mathrm{y}} \end{pmatrix} \\
&= \frac{1}{1-\rho^2}\left[\frac{(\mathrm{x}-\mu_{\mathrm{x}})^2}{\sigma_{\mathrm{x}}^2} - 2\rho\frac{(\mathrm{x}-\mu_{\mathrm{x}})(\mathrm{y}-\mu_{\mathrm{y}})}{\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}} + \frac{(\mathrm{y}-\mu_{\mathrm{y}})^2}{\sigma_{\mathrm{y}}^2}\right].
\end{aligned} \tag{2.34}$$

Finally, we can write the p.d.f. as

$$p_{\mathrm{xy}}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_{\mathrm{x}}^2} - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}} + \frac{(y-\mu_y)^2}{\sigma_{\mathrm{y}}^2}\right]\right\} \tag{2.35}$$

Now we can easily see that if x and y are uncorrelated, ρ=0, so that $p_{\mathrm{xy}}(x,y) = p_{\mathrm{x}}(x)p_{\mathrm{y}}(y)$, i.e. x and y are independent. In general, uncorrelated x and y are NOT independent, but in the special case that they are both Gaussian then they are independent. While this is easy to see here for the 2-dimensional case, Todling (1999) proves this fact for the n-dimensional case.

Clearly, Gaussian p.d.f.s have a number of nice properties that we will take advantage of, later on. If $\mathbf{x}$ is Normal or Gaussian,

- its p.d.f. is completely specified by its mean and covariance,
- if 2 Normal r.v.s are uncorrelated, they are also independent,
- A linear transformation of Normal r.v.s leads to normal r.v.s (see Exercise 4). If $\mathbf{x}$ is $\mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}}, \mathbf{R}_{\mathbf{x}})$ and $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$ then $\mathbf{y}$ is $\mathcal{N}(\boldsymbol{\mu}_{\mathbf{y}}, \mathbf{A}\mathbf{R}_{\mathbf{x}}\mathbf{A}^{\mathrm{T}})$ where $\boldsymbol{\mu}_{\mathbf{y}} = \mathbf{A}\boldsymbol{\mu}_{\mathbf{x}} + \mathbf{b}$ and $\mathbf{A}\mathbf{R}_{\mathbf{x}}\mathbf{A}^{\mathrm{T}}$ is a diagonal matrix. This means that we can find a matrix $\mathbf{A}$ that produces new r.v.s which are statistically independent.
- All marginal and conditional densities associated with variates are also Normal (see Todling (1999) ch. 2).

2.3.5 Conditional Expectation

We now extend the concept of conditional probabilities to that of conditional probability densities. If $\mathbf{x}$ and $\mathbf{y}$ are r.v.s, the conditional probability of $\mathbf{x}$ given that $\mathbf{y}$ has occurred is:

$$p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y}) = \frac{p_{\mathbf{xy}}(\boldsymbol{x},\boldsymbol{y})}{p_{\mathbf{y}}(\boldsymbol{y})}. \tag{2.36}$$

Similarly, we have that

$$p_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y}|\boldsymbol{x}) = \frac{p_{\mathbf{xy}}(\boldsymbol{x},\boldsymbol{y})}{p_{\mathbf{x}}(\boldsymbol{x})}. \tag{2.37}$$

Combining (2.36) and (2.37) gives the very important **Bayes Rule**:

$$p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y}) = \frac{p_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y}|\boldsymbol{x})p_{\mathbf{x}}(\boldsymbol{x})}{p_{\mathbf{y}}(\boldsymbol{y})}. \tag{2.38}$$

Bayes rule is very important because in data assimilation, we often know the p.d.f. of the observation error but would like to infer the p.d.f. of the model state. We can now define a conditional expectation as

$$\mathcal{E}(\mathbf{x}|\mathbf{y}) = \int_{-\infty}^{\infty} \boldsymbol{x} p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x}. \tag{2.39}$$

The marginal p.d.f. is given by

$$p_{\mathbf{x}}(\mathbf{x}) = \int_{-\infty}^{\infty} p_{\mathbf{xy}}(\boldsymbol{x},\boldsymbol{y}) d\boldsymbol{y} = \int_{-\infty}^{\infty} p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y}) p_{\mathbf{y}}(\boldsymbol{y}) d\boldsymbol{y}. \tag{2.40}$$

Note that the unconditional mean is

$$\begin{aligned}
\mathcal{E}(\mathbf{x}) &= \int_{-\infty}^{\infty} \boldsymbol{x} p_{\mathbf{x}}(\boldsymbol{x}) d\boldsymbol{x} \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \boldsymbol{x} p_{\mathbf{xy}}(\boldsymbol{x},\boldsymbol{y}) d\boldsymbol{y} d\boldsymbol{x} \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \boldsymbol{x} p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y}) p_{\mathbf{y}}(\boldsymbol{y}) d\boldsymbol{y} d\boldsymbol{x} \\
&= \int_{-\infty}^{\infty} p_{\mathbf{y}}(\boldsymbol{y}) d\boldsymbol{y} \int_{-\infty}^{\infty} \boldsymbol{x} p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x} \\
&= \int_{-\infty}^{\infty} \mathcal{E}(\mathbf{x}|\mathbf{y}) p_{\mathbf{y}}(\boldsymbol{y}) d\boldsymbol{y} \\
&= \mathcal{E}\left(\mathcal{E}(\mathbf{x}|\mathbf{y})\right).
\end{aligned} \tag{2.41}$$

The equality:

$$\mathcal{E}(\mathbf{x}) = \mathcal{E}\left(\mathcal{E}(\mathbf{x}|\mathbf{y})\right) \tag{2.42}$$

is called the **Chain Rule for conditional expectation**.

The conditional covariance matrix is given by

$$\mathbf{P}_{\mathbf{x}|\mathbf{y}} = < (\mathbf{x} - < \mathbf{x}|\mathbf{y} >)(\mathbf{x} - < \mathbf{x}|\mathbf{y} >)^{\mathrm{T}} | \mathbf{y} > .$$

Note that we have introduced a new shorthand for the expectation operator, namely, the triangular brackets. Thus,

$$\mathcal{E}(\mathbf{x}) = < \mathbf{x} > .$$

Appendix: The Gamma Function

The Gamma function is defined as

$$\Gamma(n) = \int_0^\infty e^{-x} x^{n-1} dx. \tag{2.43}$$

For integer n, it is easy to show that

$$\Gamma(n+1) = n\Gamma(n). \tag{2.44}$$

Simply evaluate (2.43) for $n+1$, then integrate by parts.

$$\begin{aligned} \Gamma(n+1) &= \int_0^\infty e^{-x} x^n dx \\ &= -x^n e^{-x}|_0^\infty + n\int_0^\infty e^{-x} x^{n-1} dx \\ &= n\Gamma(n). \end{aligned} \tag{2.45}$$

Also, by recursively substituting (2.44), one can obtain:

$$\begin{aligned} \Gamma(n+1) &= n\Gamma(n) \\ &= n(n-1)\Gamma(n-1) \\ &= n(n-1)(n-2)\Gamma(n-2) \\ &= n(n-1)(n-2)\dots(2)(1)\Gamma(1) \\ &= n(n-1)(n-2)\dots(2)(1) \\ &= n! \end{aligned} \tag{2.46}$$

Finally, note that we can directly evaluate the Γ function for non-integer values. As an example, we evaluate $\Gamma(1/2)$.

$$\Gamma(\frac{1}{2}) = \int_0^\infty e^{-x} \frac{1}{\sqrt{x}} dx \tag{2.47}$$

With the change of variables,

$$x = \frac{1}{2}y^2,$$

the above can be written as

$$\begin{aligned} \Gamma(\frac{1}{2}) &= \int_0^\infty e^{-\frac{y^2}{2}} \frac{\sqrt{2}}{y} y dy \\ &= \sqrt{2}\int_0^\infty e^{-\frac{y^2}{2}} dy \\ &= \frac{\sqrt{2}}{2}\int_{-\infty}^\infty e^{-\frac{y^2}{2}} dy \\ &= \frac{\sqrt{2}\sqrt{2\pi}}{2}\left[\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy\right] \\ &= \sqrt{\pi}. \end{aligned} \tag{2.48}$$

To obtain the third line, we used the fact that the integrand is symmetric. To obtain the 4th line, we manipulate the integrand to look like a Gaussian p.d.f. $\mathcal{N}(0,1)$ which we know integrates to 1 over the interval $[-\infty, \infty]$.

REFERENCES

1. Brown, R. G. and P. Y. C. Hwang, 1997: *Introduction to random signals and applied Kalman filtering.* John Wiley and sons, 484 pp.

2. Daley, R., 1991: *Atmospheric data analysis.* Cambridge University Press, 457 pp.

3. Maybeck, P. S., 1979: *Stochastic models, estimation and control.* Vol. 1, Academic Press, 423 pp.

4. Papoulis, A., 1965: *Probability, random variables and stochastic processes.* McGraw-Hill. 583 pp.

5. Todling, R., 1999: *Estimation theory and foundations of atmospheric data assimilation*, DAO Office Note 1999-01.