3

PROBABILITY AND EXPECTANCY

> Chance, too, which seems to rush along with slack reins, is bridled and governed by law.
>
> Anicius Manlius Severinus Boëthius[1] (~480–525 CE)

3.1 CHAPTER FOCUS

There are useful properties of certain statistical parameters of probability distributions that are the same for all probability distributions. One does not need to assume that the probability distribution is Gaussian or any other specific distribution.

These properties are very useful in Kalman filtering, and this is the essential focus of this chapter. The purpose here is to develop the essential notation and theory behind probability distributions as needed for defining and understanding Kalman filtering, which uses probability distributions defined on n-dimensional real linear spaces or manifolds. The mathematical principles involved are generally applicable to more abstract settings, but that will be the primary interest. The major subtopics are as follows:

Mathematical Foundations. It is assumed that the reader has some familiarity with the mathematical foundations of probability theory, as covered by Billingsley [1], Grinstead and Snell [2], or Papoulis [3], for example. Some of the problems

[1] Boëthius was a Roman official under the Ostrogoth King Theodoric the Great (454–526 CE). The quote is from *De Consolatione Philosophiae,* written in 574 CE, when Boëtius was in prison awaiting trial and eventual execution on charges of treason.

Kalman Filtering: Theory and Practice Using MATLAB®, Fourth Edition.
Mohinder S. Grewal and Angus P. Andrews.

at the end of the chapter will test your background knowledge. If you have trouble with them, perhaps you should refresh your knowledge by consulting the references at the end of the chapter or equivalent sources. These sources also define the underlying measure theory, in terms of what is meant by the *union* or *intersection* of measurable sets. The treatment of these underlying concepts in this chapter is heuristic and brief.

Probability Density Functions. It makes all this a bit more transparent if we identify probability distributions on real n-dimensional spaces in terms of their *probability density functions* and then use the ordinary calculus to demonstrate the essential properties of the resulting expectancy operators.

Expectancy, Moments, and Optimal Estimates. The key operator needed for developing the essential formulas is what is called the *expectancy operator*, which turns out to be a *linear functional*—a very powerful mathematical tool for our purposes. It is used for defining the *moments* of probability distributions, and the first and second moments of probability distributions on n-dimensional real space $\mathfrak{R}^n$ can be identified with the essential variables used in Kalman filtering. Furthermore, so long as all transformations of the variables involved are linear, their effects on those moments are expressible in simple formulas that are not dependent on any particulars of the underlying probability distributions beyond their moments. These formulas lead directly to the identification of the estimator with the least-mean-squared error—independent of the specific probability distribution involved. This unshackles least-mean-squared estimation from dependency on any particular error probability distribution.

As a reality check, several probability distributions with different probability density functions are used to demonstrate that the important results do not depend on what particular probability density functions is used—so long as it has definable first and second moments.

Nonlinear Effects. In preparation for Chapter 8 (nonlinear extensions), effects of nonlinear transformations on the expected values (means) and expected squared deviation from the mean (covariances) of probability distributions are also modeled and discussed.

3.2 FOUNDATIONS OF PROBABILITY THEORY

Probabilities are a way of quantifying uncertainty. Probability theory is about mathematical models for doing just that.

Its mathematical foundations started with models for gambling using dice or playing cards, in which the number of possible outcomes is generally manageable with a bit of mental arithmetic. Gamblers were primarily concerned with the odds of different outcomes and the expected long-term winnings from different betting strategies. These were often determinable with a pencil and a paper, given mathematical models appropriate to the game.

But uncertainties can also include such things as the miss distance of a projectile, in which case the possible outcomes are modeled as real numbers. To handle cases such as this, mathematicians would eventually resort to *probability measures*, which characterize the outcomes of the modeled activities in terms of nonnegative numbers. Here, we will not concern ourselves much with measure theory, except to give assurances that it provides the essential mathematical underpinnings of probability theory, and to give some general notion of what it is and what it does. Readers seeking more formal coverage can find it in References [1–3].

3.2.1 Measure Theory

Probability theory changed in the early twentieth century with the development of *measure theory* as part of the foundations of the integral calculus. The rebuilding of probability on the foundation of measure theory had many contributors [4], with a major role played by Andrey N. Kolmogorov (1903–1987) [5]. The theory may not be particularly intuitive in nature, but it does serve as a foundation for integration and probability theory. Its treatment here will be rather cursory.

Measures are nonnegative functions defined on the so-called *sigma algebras* of "measurable sets" in the domain of a potentially integrable function. Measures are not functions in the ordinary sense, because they do not assign a specific value to each point, but to each "measurable set" of points. The base measurable sets for the ordinary (Riemann or Riemann–Stieltjes) integral of the calculus, for example, are built up from finite intervals, using the length of the interval as its measure. The term *sigma-algebra* refers to the fact that the measurable sets obey certain laws about the measures of the set unions and intersections of measurable sets.[2]

Measure theory does not necessarily change the way calculus is done. It mostly shores up its mathematical foundations and—in the process—defines which functions are integrable.

When we write an integral as

$$\int f(x)\, dx,$$

the $\int$ comes from a stylized "S," used as shorthand for "Sum," and the "dx" represents the measure used in defining the integral. Every real function defined somewhere on the real line has a *domain* (set of real values for which the function is defined) and *range* (set of real values the function assigns to numbers in its domain). The measure used for defining integration essentially defines the resulting classes of integrable functions, but the resulting integral calculus is otherwise much the same.

3.2.2 Probability Measures

The distinguishing feature of probability measures is that the probability measure of the union of all measurable sets is always equal to 1 (unity). That is, the probability

[2] Set theory defines an *algebra* of sets, under the operations of set union and set intersection. The "sigma" refers to the additive (arithmetic) properties of measures of set unions and intersections.

that a random sample being in the domain on which it is defined is 1 (one). Just to hedge a bit, probability theorists might call that outcome "almost certain," and only claim such an outcome would "almost surely" happen.

On the other end of things, the probability that an arrow aimed at a target would strike exactly in the center of the bullseye would be zero (except in fiction). There are just too many other places where the arrow could strike.

The value assigned to a measurable set of points by a probability measure is the probability that a random sampling from the probability distribution would lie in that set of points.

There is not much difference between the mathematical properties of probability measures and ordinary measures, other than the fact that probability integrals of the entire domain must equal 1. Integrals with respect to a probability measure p might use the notation dp or $dp(x)$ in place of the Riemann measure dx, but otherwise integrals involving probability measures look and act much the same as ordinary integrals.

3.2.3 Probability Distributions

A *probability distribution* is defined by a probability measure on the measurable subsets of a set of points $\mathcal{S}$. In general, $\mathcal{S}$ could be any abstract measurable set, but the focus here will be exclusively on n-dimensional real spaces, or on multiply connected sets such circles or surfaces of three-spheres in four-dimensional real spaces (used for representing attitude).

We might, for example, define a probability measure over the surface of Earth (a two-dimensional real manifold embedded in a three-dimensional universe) modeling the likelihood that some object approaching Earth from space would strike there. In this case, however, the integral of that probability measure over the entire surface could be <1, with the leftover probability assigned to the rest of the universe to represent the case that the object would miss Earth altogether.

3.2.3.1 Some Definitions

Probability Spaces Formally, a probability space is defined by a triplet $\{\mathcal{S}, \mathcal{A}, \mathcal{P}\}$, where

$\mathcal{S}$ is called the *set of all possible outcomes* of a random event.

$\mathcal{A}$ is a *sigma-algebra* of subsets of $\mathcal{S}$, meaning that if $\mathcal{A}$ includes sets $A \subset \mathcal{S}$ and $B \subset \mathcal{S}$, then it includes their set union $A \cup B$ and set intersection $A \cap B$. The sigma-algebra $\mathcal{A}$ must also contain $\mathcal{S}$ and its "$\mathcal{S}$-complement" $\mathcal{S} - \mathcal{S} = \emptyset$, the *empty set.*

$\mathcal{P}$ is a *probability measure,* which assigns a nonnegative value to all sets $A \in \mathcal{A}$ and $B \in \mathcal{A}$ such that

- $\mathcal{P}(\emptyset) = 0$.

- $\mathcal{P}(\mathcal{S}) = 1$.
- $\mathcal{P}(A \cup B) = \mathcal{P}(A) + \mathcal{P}(B) - \mathcal{P}(A \cap B)$.

We will not be using these definitions and properties, but you should be aware that they are part of the foundations of probability theory.

Random Variables (RV). The alternative terms *variate*, *random variable* (RV), or *stochastic variable* are used to denote possible *outcomes* of a draw from a probability distribution, such as drawing numbered balls in a lottery. These terms do not denote a specific outcome, but the ensemble of possible outcomes and their associated probabilities. If the number of possible outcomes is finite, then each outcome may have a specific probability. However, for probability spaces in which the set of possible outcomes $\mathcal{S}$ is a real n-dimensional vector space or manifold, the probability of drawing a specific point $x \in \mathcal{S}$ is more likely to be zero. This is a rather counter-intuitive concept, but it is a consequence of the vastness of the real numbers.

Realizations of Variates. A specific point $x \in S$ would be called a *realization* of a variate, random variable, stochastic variable, or outcome.

Notation. We will observe the common practice of using an uppercase letter to designate a *variate* X, representing the ensemble of values that might be drawn from a probability distribution and to use the corresponding lowercase letter x to denote a *realization* of that draw. The notation $x \in X$ would then indicate that x is a realization of the variate X.

3.2.4 Probability Density Functions

We are perhaps more familiar with having a probability distribution defined in terms of a nonnegative *probability density function* $p(x) \geq 0$ defined for x in n-dimensional real space $\Re^n$. In this case, the fact that it is a probability measure means that its integral

$$\int_{\Re^n} p(x)\ dx = 1. \tag{3.1}$$

The equivalent probability measure of any measurable subset A of $\Re^n$ would then be defined as the integral of the probability density function p over that subset of $\Re^n$, which is also the probability that a randomly chosen value of x would lie in A:

$$P(x \in A) = \int_A p(x)\ dx. \tag{3.2}$$

In this case, the equivalent probability measure is essentially $p(x)\ dx$, where dx represents the Riemann or Riemann–Stieltjes measure—or any suitable measure, for that matter.

Example 3.1 (Gaussian Probability Density) The probability distribution of the average number of occurrences of a "one" (•) in N successive tosses of a die has what is called a "binomial" distribution. As $N \to \infty$, this tends toward a particular type of distribution called a "*Gaussian*" or "*normal*" distribution. The Gaussian/normal distribution[3] is the limit of many other distributions, and it is common to many models for random phenomena. It is commonly used in stochastic system models for the distributions of random variables.

The *multivariate* Gaussian probability density function for a column vector x of n components has the form

$$p(x) = \frac{1}{\sqrt{2\pi \det P_{xx}}} \exp \left(-\frac{1}{2}\left(x - \mu_x\right)^T P_{xx}^{-1}(x - \mu_x)\right)$$

$$x \in \mathcal{N}(\mu_{\mathrm{x}}, \mathcal{P}_{\mathrm{xx}}), \tag{3.3}$$

with parameters

μ_x, the *mean* of the distribution, a column vector with n components, and

P_{xx}, the *covariance* of the distribution, an $n \times n$ symmetric positive-definite matrix.

The "$\mathcal{N}$" notation stands for "normal." However, so many things in mathematics are called "normal," it is called "Gaussian" here to avoid ambiguity.

General Probability Integrals A probability measure defines which functions are integrable under that probability measure. In that case, the integral of an integrable function $f(x)$ over a measurable set A could be denoted as

$$\int_A f(x)dp(x),$$

where $dp(x)$ denotes the probability measure being used. This notation allows for the possibility that the probability measure cannot be represented in terms of a probability density function.

Example 3.2 (The Dirac Delta Distribution) Not all legitimate probability distributions can be defined by density functions. The Dirac δ "function," for example, is not a true function, but it can be defined as the probability measure which assigns the value 1 to any measurable set containing a specified point (the mean of the resulting probability distribution). This Dirac probability measure has the property that its variance is zero, implying that there is no uncertainty about the value of its variate. (It is the mean of the distribution, almost surely.) However, the probability distributions of interest in Kalman filtering are not very likely to have zero variance, and problems with variances too close to zero can be ill-conditioned for Kalman filtering implementations in finite-precision arithmetic.

[3] It's probability density function has the property that its Fourier transform is also a Gaussian probability density function. Physicist Gabriel Lippman (1845–1921) is credited with the observation that "mathematicians think it [the normal distribution] is a law of nature and physicists are convinced that it is a mathematical theorem."

3.2.5 Cumulative Probability Functions

For any probability measure $dp(x)$ defined on the real line, the function $P(x)$ defined as

$$P(x) \overset{\text{def}}{=} \int_{-\infty}^{x} d(x) \tag{3.4}$$

is called the *cumulative probability function* (or just *probability function*) of the probability measure.

If—unlike the Dirac distribution—the probability distribution in question has a density function

$$p(x)\ dx = dp(x),$$

then its cumulative probability function

$$P(x) \overset{\text{def}}{=} \int_{-\infty}^{x} p(\chi)\ d\chi. \tag{3.5}$$

Example 3.3 (Gaussian Cumulative Probability Function) The probability function for the univariate Gaussian distribution with mean μ and variance σ^2 is defined as

$$P_{\mathcal{N}}(x,\ \mu,\ \sigma) \overset{\text{def}}{=} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \exp\left(-\frac{(\chi-\mu)^2}{2\sigma^2}\right)\ d\chi \tag{3.6}$$

$$= \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)\right], \tag{3.7}$$

where erf is an analytic function called the “error function” (callable in MATLAB® as `erf`).

3.3 EXPECTANCY

3.3.1 Linear Functionals

Linear functionals are powerful devices in mathematical analysis, and one of them—the expectancy operator—is especially useful in estimation theory.

Any mapping $\mathcal{F}$ from real functions $f(\cdot)$ to real numbers is called a *functional.*

It is called a *linear* functional if, for any two such functions $f(\cdot)$ and $g(\cdot)$ and for any two real numbers a and b,

$$\mathcal{F}[a\, f(\cdot) + b\ g(\cdot)] = a\ \mathcal{F}[f(\cdot)] + b\ \mathcal{F}[b\ g(\cdot)]. \tag{3.8}$$

Linear functionals play a major role in the definition of *generalized functions* such as the Dirac δ-function and probability measures with no corresponding density functions.

3.3.2 Expectancy Operators

3.3.2.1 Expected Values and Probability Densities The *expected value* of any function $f(x)$ of a variate x from a probability distribution with density function $p(\cdot)$ is defined by the operator

$$\underset{x}{\mathrm{E}}\langle f(x)\rangle \overset{\text{def}}{=} \int f(x)p(x)\, dx \tag{3.9}$$

provided the integral exists. If so, the result is a real number and the *expectancy operator* is defined as the mapping of functions to their expected values. We could also represent the operator as

$$\underset{x\in X}{\mathrm{E}}\langle\cdot\rangle$$

to indicate the probability distribution of the variate under consideration (X), but there is not much room for it under the E.

The expectancy operator is also a *linear functional* because, for any real numbers a and b and for any integrable functions $f(\cdot)$ and $g(\cdot)$ defined on the domain of the probability density function $p(\cdot)$,

$$\underset{x}{\mathrm{E}}\langle af(x)+bg(x)\rangle = \int [af(x)+bg(x)]\, p(x)\, dx \tag{3.10}$$

$$= \int a\, f(x)\, p(x)\, dx + \int b\, g(x)\, p(x)\, dx \tag{3.11}$$

$$= a \int f(x)\, p(x)\, dx + b \int g(x)\, p(x)\, dx \tag{3.12}$$

$$= a\ \underset{x}{\mathrm{E}}\langle f(x)\rangle + b\ \underset{x}{\mathrm{E}}\langle g(x)\rangle. \tag{3.13}$$

This result also applies to cases where x and the functions $f(\cdot)$ and $g(\cdot)$ have vector values.

3.3.2.2 Expected Values and Probability Measures The expectancy operator has been defined in terms of a probability density function $p(x),\ x \in X$. However, it can as well be defined in terms of the probability measure of X, without reference to a probability density function.

A probability measure $\wp(\cdot)$ assigns a nonnegative real value to every measurable subset $S \subseteq \mathcal{S}$ in a sigma-algebra of subsets of a parent set $\mathcal{S}$. The *characteristic function* of a measurable set S is the function

$$f_S\ (x) \overset{\text{def}}{=} \begin{cases} 1, & x \in S \\ 0, & x \notin S \end{cases} \tag{3.14}$$

and the associated expectancy operator for that probability measure assigns the value

$$\underset{x}{\mathrm{E}}\langle f_S\ (x)\rangle = \wp\ (S), \tag{3.15}$$

where $\wp\ (S)$ is the probability measure of S.

Example 3.4 (Expected Values and Sampling Functionals) The basic Dirac measure assigns the value 1 (one) to all real measurable sets containing the value 0 (zero). The associated expectancy operator, then, would assign the value

$$\underset{x}{\mathrm{E}}\langle f(x)\rangle = f(0)$$

for every Dirac-measurable function f. By subtracting a more desirable sampling argument x_0, this becomes the sampling "function" (functional) for the value of the function f at x_0:

$$\underset{x}{\mathrm{E}}\langle f(x - x_0)\rangle = f(x_0).$$

Note that this is still a linear functional.

3.3.3 Moments of Distributions

If they exist, the expected values of powers of variates of a scalar probability distribution,

$$\underset{x}{\mathrm{E}}\langle x^N\rangle,$$

are called the "*raw*" *moments* of the distribution. Not all distributions have finite raw moments, however.

Example 3.5 (Cauchy Distribution) The Cauchy distribution[4] is used in risk analysis. In quantum mechanics, it is called the *Lorentz distribution* and used for modeling energy distributions of unstable states.

It has a parametric probability density function of the form

$$p_{\mathrm{Cauchy}}(x,\ m,\ \gamma) \overset{\mathrm{def}}{=} \frac{\gamma/\pi}{\gamma^2 + (x - m)^2}, \tag{3.16}$$

defined for $-\infty < x < +\infty$. The parameter m in this case is both the *mode* (maximum probability density) and *median* (value at which the cumulative probability equals $1/2$) of the distribution. The additional positive parameter γ is called a *scaling parameter*. It behaves somewhat like the standard deviation of the distribution, in that the distribution becomes more spread out as γ increases.

However, the Cauchy distribution has no finite mean (first moment) and variance (second central moment).

If they do exist, these moments are constant parameters of the distribution in question, and their values can provide an alternative definition of the distribution.

The notation ${}^{[N]}\mu_x$ will be used to denote the Nth raw moment of a distribution with the variate X—if the moment in question exists.

[4] Named after French mathematician Baron Augustin-Louis Cauchy (1789–1857).

There will always be a *zeroth-order moment*, the scalar

$$ {}^{[0]}\mu_x \overset{\text{def}}{=} \underset{x}{\mathrm{E}}\langle x^0\rangle = \underset{x}{\mathrm{E}}\langle 1\rangle = \int p_x(x)\,dx = 1, \tag{3.17}$$

because the underlying measure of integration is a probability measure. The zeroth-order moment is always a scalar, independent of the dimension of the variate X involved.

3.3.3.1 Means If it does exists, the first-order moment of the n-dimensional variate X of a multivariate probability distribution will be a vector of the same dimension,

$$ {}^{[1]}\mu_x \overset{\text{def}}{=} \underset{x}{\mathrm{E}}\langle x\rangle \overset{\text{def}}{=} \begin{bmatrix} \underset{x}{\mathrm{E}}\langle x_1\rangle \\ \underset{x}{\mathrm{E}}\langle x_2\rangle \\ \underset{x}{\mathrm{E}}\langle x_3\rangle \\ \vdots \\ \underset{x}{\mathrm{E}}\langle x_n\rangle \end{bmatrix}, \tag{3.18}$$

called the *mean* of the distribution of X.

3.3.3.2 Central Moments After the mean, the moment of first order ($N = 1$), there are two choices for defining the higher order moments, and the dimensionality of the data structures representing moments of order $N > 1$ depends on N and the dimensionality of the underlying variate X.

Besides the *raw moment* used for defining the mean, there are *central moments* defining the moments "about the mean." For a scalar variate X, the moments of order $N > 1$ would be defined as

$$ {}^{[N]}\mu_x \overset{\text{def}}{=} E_x\langle x^N\rangle \qquad \text{(raw moment)} \tag{3.19}$$

$$ {}^{[N]}\sigma_x \overset{\text{def}}{=} E_x\langle (x - {}^{[1]}\mu_x)^N\rangle \qquad \text{(central moment).} \tag{3.20}$$

Here, the notation ${}^{[N]}\sigma_x$ is used to distinguish the Nth central moment from the raw moment, ${}^{[N]}\mu_x$.

3.3.3.3 Covariance Matrices After the mean ($N = 1$), the next higher order *central moment* ($N = 2$) of an n-dimensional vector variate X is an $n \times n$ matrix

$$ P_{xx} \overset{\text{def}}{=} {}^{[2]}\sigma_x \tag{3.21}$$

$$ = \underset{x}{\mathrm{E}}\langle (x - {}^{[1]}\mu_x)(x - {}^{[1]}\mu_x)^{\mathrm{T}}\rangle \tag{3.22}$$

$$= \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1n} \\ p_{21} & p_{22} & p_{23} & \cdots & p_{2n} \\ p_{31} & p_{32} & p_{33} & \cdots & p_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & p_{n3} & \cdots & p_{nn} \end{bmatrix} \tag{3.23}$$

$$p_{ij} \overset{\text{def}}{=} E_x\langle (x_i - {}^{[1]}\mu_{x\ i})(x_j - {}^{[1]}\mu_{x\ j})\rangle. \tag{3.24}$$

The equivalent second-order *raw moment* is also an $n \times n$ matrix,

$${}^{[2]}\mu_x = \underset{x}{\mathrm{E}}\langle xx^{\mathrm{T}}\rangle \tag{3.25}$$

$$= \begin{bmatrix} {}^{[2]}\mu_{x\ 11} & {}^{[2]}\mu_{x\ 12} & {}^{[2]}\mu_{x\ 13} & \cdots & {}^{[2]}\mu_{x\ 1n} \\ {}^{[2]}\mu_{x\ 21} & {}^{[2]}\mu_{x\ 22} & {}^{[2]}\mu_{x\ 23} & \cdots & {}^{[2]}\mu_{x\ 2n} \\ {}^{[2]}\mu_{x\ 31} & {}^{[2]}\mu_{x\ 32} & {}^{[2]}\mu_{x\ 33} & \cdots & {}^{[2]}\mu_{x\ 3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ {}^{[2]}\mu_{x\ n1} & {}^{[2]}\mu_{x\ n2} & {}^{[2]}\mu_{x\ n3} & \cdots & {}^{[2]}\mu_{x\ nn} \end{bmatrix} \tag{3.26}$$

$${}^{[2]}\mu_{x\ ij} \overset{\text{def}}{=} E_x\langle x_i x_j\rangle. \tag{3.27}$$

3.3.3.4 Moments of Order N > 2 For scalar variates X, these are scalars defined by Equations 3.19 (raw moments) and 3.20 (central moments).

When the variates X are n-dimensional vectors, it is a bit more complicated. The general Nth-order moment will be a data structure with dimensions

$$\underbrace{n \times n \times n \times \cdots \times n.}_{N \text{ times}}$$

Any Nth-order raw moment ${}^{[N]}\mu_x$ is a data structure indexed over N subscripts $i_1, i_2, i_3, \ldots, i_N$, with the jth subscript $1 \leq i_j \leq n$ and the corresponding data structure elements

$${}^{[N]}\mu_{xi_1,\ i_2,\ i_3,\ \ldots,\ i_N} \overset{\text{def}}{=} \underset{x}{\mathrm{E}}\langle x_{i_1} \times x_{i_2} \times x_{i_3} \times \cdots \times x_{i_N}\rangle, \tag{3.28}$$

The corresponding Nth-order central moment ${}^{[N]}\sigma_x$ would have $(x_{i_j} - {}^{[1]}\mu_{x\ i_j})$ in place of x_{i_j} in Equation 3.28.

Example 3.6 (Moments of Univariate Gaussian Distributions) If the distribution in question is univariate Gaussian with mean μ and variance σ^2, the higher order central moments are all determinable from μ and σ as

$$\sigma_k = \begin{cases} 0, & k \text{ odd}, \\ \sigma^k \ (k-1)\,!!, & k \text{ even}, \end{cases} \tag{3.29}$$

$$(k-1)!! \overset{\text{def}}{=} 1 \cdot 3 \cdot 5 \cdot 7 \cdot \cdots \cdot (k-1), \tag{3.30}$$

the so-called "double factorial" for odd numbers.

The higher order raw moments are also determinable from μ and σ, but the relationships are a bit more complex:

$$\mu_k = \begin{cases} \pi^{-1/2} 2^{(k+1)/2} \mu \sigma^{k-1} \Gamma\left(\dfrac{k+1}{2}\right) K\left(\dfrac{1-k}{2}, \dfrac{1}{2}, \dfrac{-\mu^2}{2\sigma^2}\right) & k \text{ odd}, \\ \pi^{-1/2} 2^{k/2} \sigma^k \Gamma\left(\dfrac{k+3}{2}\right) K\left(\dfrac{-k}{2}, \dfrac{3}{2}, \dfrac{-\mu^2}{2\sigma^2}\right) & k \text{ even}, \end{cases} \tag{3.31}$$

where the Gamma function is the positive-real-valued extension of the factorial function,

$$\Gamma(z) \overset{\text{def}}{=} \int_0^{+\infty} s^{z-1} e^{-s} \, ds, \tag{3.32}$$

Kummer's confluent hypergeometric function ${}_1F_1$

$$K(a,\ b,\ c) \overset{\text{def}}{=} \sum_{k=0}^{+\infty} \frac{(a)_k \ c^k}{(b)_k \ k!}, \tag{3.33}$$

and the *Pochammer notation*

$$(y)_k \overset{\text{def}}{=} \frac{\Gamma(y+k)}{\Gamma(y)}. \tag{3.34}$$

Evaluations of Equations 3.31 and 3.29 should produce the values shown in Table 3.1. Unlike the case with raw moments, all odd-order central moments are zero, and all even-order central moments depend only on the second-order central moment.

3.3.3.5 Cross-Covariance

Let p_{ij} be the element in the ith row and jth of a covariance matrix P_{xx} of a vector variate X,

$$p_{ij} = \mathop{\mathrm{E}}_{x}\langle (x_i - \mu_i)(x_j - \mu_j) \rangle. \tag{3.35}$$

p_{ij} is called the *cross-covariance*[5] of the ith and jth components of X.

[5] Strictly speaking, if $i = j$, it is just the *variance* of the ith $= j$th component of X. Only if $i \neq j$, is it called the *cross-covariance* of the scalar variates X_1 and X_j.

TABLE 3.1 Scalar Gaussian Moments

Order N	Central Moment	Raw Moment
0	1	0
1	0	μ
2	σ^2	$\mu^2 + \sigma^2$
3	0	$\mu^3 + 3\mu\sigma^2$
4	$3\sigma^4$	$\mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$
5	0	$\mu^5 + 10\mu^3\sigma^2 + 15\mu\sigma^4$
6	$15\sigma^6$	$\mu^6 + 15\mu^4\sigma^2 + 45\mu^2\sigma^4 + 15\sigma^6$
7	0	$\mu^7 + 21\mu^5\sigma^2 + 105\mu^3\sigma^4 + 105\mu\sigma^6$
8	$105\sigma^8$	$\mu^8 + 28\mu^6\sigma^2 + 210\mu^4\sigma^4 + 420\mu^2\sigma^6 + 105\sigma^8$
9	0	$\mu^9 + 36\mu^7\sigma^2 + 378\mu^5\sigma^4 + 1260\mu^3\sigma^6 + 945\mu\sigma^8$
10	$945\sigma^{10}$	$\mu^{10} + 45\mu^8\sigma^2 + 630\mu^6\sigma^4 + 3150\mu^4\sigma^6 + 4725\mu^2\sigma^8 + 945\sigma^{10}$

More generally, if X_a and X_b represent nonoverlapping subvectors of the vector variate X, then the submatrix

$$P_{ab} = \underset{x}{\mathrm{E}}\langle x_a x_b^{\mathrm{T}}\rangle \tag{3.36}$$

of P_{xx} represents the cross-covariance of X_a and X_b.

For example, if X is partitioned into the subvectors

$$x = \begin{bmatrix} X_a \\ X_b \end{bmatrix}, \tag{3.37}$$

then the corresponding covariance matrix can be partitioned as

$$P_{xx} = \begin{bmatrix} P_{x_a x_a} & P_{x_a x_b} \\ P_{x_b x_a} & P_{x_b x_b} \end{bmatrix}, \tag{3.38}$$

where

$P_{x_a x_a}$ is the covariance of X_a,
$P_{x_b x_b}$ is the covariance of X_b,
$P_{x_a x_b}$ is the cross-covariance of X_a and X_b, and
$P_{x_b x_a}$ is the cross-covariance of X_b and X_a.

In general, however,

$$P_{x_a x_b} = P_{x_b x_a}^{\mathrm{T}}. \tag{3.39}$$

3.3.3.6 Correlation Coefficients For any term p_{ij} in the *i*th row and *j*th column of a covariance matrix P_{xx}, the ratio

$$\rho_{ij} \overset{\text{def}}{=} \frac{p_{ij}}{\sqrt{p_{ii}p_{jj}}} \tag{3.40}$$

is called the *correlation coefficient* between the *i*th and *j*th components of X. It has the following properties:

1. $\rho_{ii} = 1$ for all i.
2. $\rho_{ij} = \rho_{ji}$.
3. $-1 \leq \rho_{ij} \leq 1$ for all i and j.

The first two of these assertions are consequences of the definition of ρ_{ij} in Equation 3.40. The last assertion is a consequence of Hölder's inequality, originally defined for vectors [6], but generalized to functions.

Given the values of ρ_{ij} for a covariance matrix P_{xx}, it can be decomposed as a matrix product

$$P_{xx} = \text{diag}(\sigma)C_\rho \text{ diag } (\sigma) \tag{3.41}$$

$$\text{diag } (\sigma) \overset{\text{def}}{=} \begin{bmatrix} \sigma_1 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_n \end{bmatrix} \tag{3.42}$$

$$\sigma_i \overset{\text{def}}{=} \sqrt{p_{ii}}, i = 1,\ 2,\ 3,\ \ldots\ ,\ n \tag{3.43}$$

$$C_\rho \overset{\text{def}}{=} \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2n} \\ \rho_{31} & \rho_{32} & 1 & \cdots & \rho_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \rho_{n3} & \cdots & 1 \end{bmatrix}. \tag{3.44}$$

3.3.3.7 Statistical Independence and Correlation Two variates X and Y are said to be *statistically independent* if their *joint probability density function*

$$p(x, y) = p(x)\ p(y), \tag{3.45}$$

where $p(x)$ and $p(y)$ are the product of the individual probability densities of X and Y, respectively.

For example, if the ith and jth components of a vector variate X are statistically independent, then

$$p(x_i, x_j) = p(x_i)\, p(x_j), \tag{3.46}$$

and their cross-covariance

$$\begin{aligned}
p_{ij} &\overset{\text{def}}{=} E_x\langle (x_i - \mu_i)(x_j - \mu_j)\rangle && (3.47)\\
&\overset{\text{def}}{=} \int_{x_i}\int_{x_j} (x_i - \mu_i)(x_j - \mu_j)\, p(x_i, x_j)\, dx_i\, dx_j && (3.48)\\
&= \left[\int_{x_i} \left(x_i - \mu_i\right)\ p(x_i)\, dx_i\right] \left[\int_{x_j} \left(x_j - \mu_j\right) p(x_j)\, dx_j\right] && (3.49)\\
&= [\mu_i - \mu_i]\,[\mu_j - \mu_j] && (3.50)\\
&= 0. && (3.51)
\end{aligned}$$

That is, statistical independence of components of a vector variate X is equivalent to having zero cross-covariance—or, equivalently, zero correlation coefficient.

The same applies to statistical independence of nonoverlapping subvectors $X_a,\ X_b$ of a vector variate X. In that case, the corresponding submatrix $P_{x_a\ x_b}$ of the covariance matrix P_{xx} will be a zero matrix.

For example, if the vector variate X is partitioned into the statistically independent sub-vector variates X_a and X_b such that the sample vectors

$$x = \begin{bmatrix} x_a \\ x_b \end{bmatrix}, \tag{3.52}$$

then the corresponding covariance matrix can be partitioned as

$$P_{xx} = \begin{bmatrix} P_{x_a x_a} & 0 \\ 0 & P_{x_b x_b} \end{bmatrix}, \tag{3.53}$$

where $P_{x_a x_a}$ is the covariance of X_a and $P_{x_b x_b}$ is the covariance of X_b.

3.4 LEAST-MEAN-SQUARE ESTIMATE (LMSE)

The word *estimate* comes from the Latin verb *aestimare,* meaning “to place a value on (something).” In the present context, that “value” will be a real number or real vector and the “something” will be a probability distribution defined over an n-dimensional real space by its mean and covariance.

Among the more profound discoveries in probability theory are the facts that

1. The mean of a probability distribution defined on $\Re^n$ is also the estimated value that achieves the least-mean-squared estimation error.
2. The *trace* (sum of diagonal elements) of the associated covariance matrix equals that least (or minimum) mean-squared estimation error.

3.4.1 Squared Estimation Error

An estimate has also been called *an educated guess*. Before an archer releases an arrow aimed at a target, her or his best estimate of where it will end up is in the center of the target—or as close thereto as possible.[6] The associated "estimation error" in this case is the miss distance, which is the square root of the sum of the squares of the arrow's horizontal and vertical displacements from the exact center of the target. Minimizing the miss distance in this case is equivalent to minimizing the sum of the squares of the miss vector components. In practice, the likelihood of making that error zero all the time is negligible, so the long-term goal of most archers is to minimize the expected squared miss distance.

By using the expectancy operator, that same goal can be put in terms of probability distributions in a way that defines the optimal estimate x̂ of a value drawn from that probability distribution in terms that depend only on the first moment of the distribution, but not otherwise on the shape of the associated probability density function.

3.4.2 Minimization

Optimal estimates are defined by some optimality criterion. In the case that criterion is the *expected* (i.e., *mean*) squared estimation error, the optimal estimate is the one which minimizes the mean-square estimation error. In that case, it is called the *minimum-mean-squared estimate* (*MMSE*), or *least-mean-squared estimate* (*LMSE*).

For any estimate $\hat{x}$ for the value of x in a probability distribution over n-dimensional real space with first- and second-order moments, the mean-squared estimation error can be defined as

$$\varepsilon^2(\hat{x}) = \underset{x}{\mathrm{E}}\langle |\hat{x} - x|^2 \rangle \tag{3.54}$$

$$= \underset{x}{\mathrm{E}}\langle (\hat{x} - x)^{\mathrm{T}}(\hat{x} - x) \rangle \tag{3.55}$$

$$= \underset{x}{\mathrm{E}}\langle \hat{x}^{\mathrm{T}}\hat{x} - 2\hat{x}^{\mathrm{T}}x + x^{\mathrm{T}}x \rangle \tag{3.56}$$

$$= \underset{x}{\mathrm{E}}\langle \hat{x}^{\mathrm{T}}\hat{x} \rangle - 2\,\underset{x}{\mathrm{E}}\langle \hat{x}^{\mathrm{T}}x \rangle + \underset{x}{\mathrm{E}}\langle x^{\mathrm{T}}x \rangle \tag{3.57}$$

[6] The Greek word for this guessing/aiming process is "στόχοζ," which gives us the modern word *stochastic*.

$$= \mathop{\mathrm{E}}_{x}\left\langle \sum_j \hat{x}_j^2 \right\rangle - 2 \mathop{\mathrm{E}}_{x}\left\langle \sum_j \hat{x}_j x_j \right\rangle + \mathop{\mathrm{E}}_{x}\left\langle \sum_j x_j^2 \right\rangle \tag{3.58}$$

$$= |\hat{x}|^2 - 2\hat{x}^{\mathrm{T}} \mathop{\mathrm{E}}_{x}\langle x\rangle + \mathop{\mathrm{E}}_{x}\langle |x|^2\rangle. \tag{3.59}$$

This mean-squared error $\varepsilon^2(\hat{x}) \to +\infty$ as any component $\hat{x}_j \to \pm\infty$ and achieves its minimum $\hat{x}_{\mathrm{LMSE}}$ where

$$0 = \frac{\partial \varepsilon^2(\hat{x})}{\partial \hat{x}} \tag{3.60}$$

$$= 2\hat{x} - 2 \mathop{\mathrm{E}}_{x}\langle x\rangle, \tag{3.61}$$

which can be solved for the estimate $\hat{x}_{\mathrm{LMSE}}$ with the least-mean-squared error,

$$\hat{x}_{\mathrm{LMSE}} = \mathop{\mathrm{E}}_{x}\langle x\rangle. \tag{3.62}$$

That is, the LMSE of x is the mean of the distribution—independent of the probability distribution involved.

3.4.3 Least-Mean-Squared Estimation Error

The covariance of least-mean-squared estimation error is the second central moment of the probability distribution, also known as the *covariance of the probability distribution*,

$$\mathop{\mathrm{E}}_{x}\langle (\hat{x}_{\mathrm{LMSE}} - x)(\hat{x}_{\mathrm{LMSE}} - x)^{\mathrm{T}}\rangle = P_{xx}. \tag{3.63}$$

Equation 3.55 for the squared estimation error, evaluated at $\hat{x} = \mu$, is

$$\varepsilon^2(\mu) = \sum_j \mathop{\mathrm{E}}_{x}\langle (x - \mu)^2\rangle \tag{3.64}$$

$$= \sum_j p_{jj} \tag{3.65}$$

$$= \mathrm{tr}\ (P_{xx}), \tag{3.66}$$

the matrix *trace* (sum of diagonal elements) of P_{xx}.

That is, the mean μ and covariance P_{xx} of any probability distribution characterize the LMSE, in terms of its value and its mean-squared error.

Example 3.7 (Evaluation Using Five Different Distributions) This example computes and plots the mean-square estimation error as a function of the estimated value for five different probability density functions, all defined on the real line with the same means and variances, as plotted in Figure 3.1. These include the

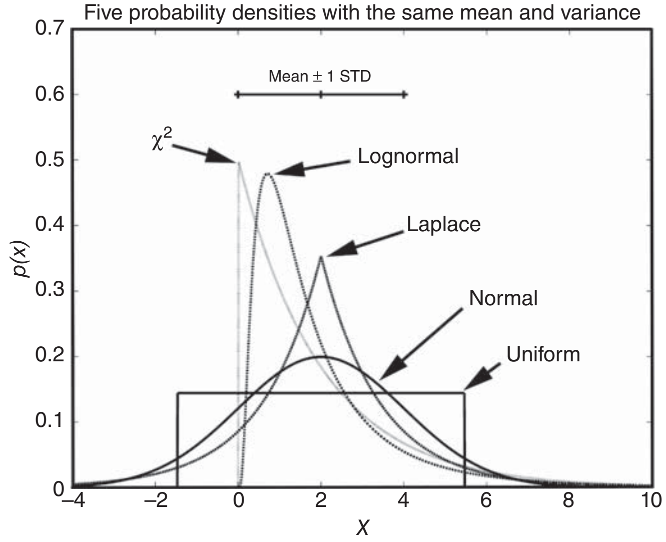

Figure 3.1 Five probability density functions with the same means and variances.

following:

1. A uniform distribution with mean equal to 2 and variance equal to 4. This has zero probability off a fixed interval on the real line. It is used to represent quantization errors in digitized signals, for example.
2. A Gaussian (or normal) distribution with mean equal to 2 and variance equal to 4. This has positive probability everywhere on the real line. The Gaussian distribution is perhaps overused a bit.
3. A Laplace distribution with mean equal to 2 and variance equal to 4. This also has positive probability everywhere on the real line but is different from the Gaussian distribution.
4. A lognormal distribution with mean equal to 2 and variance equal to 4. This is nonzero only for nonnegative input values, and has a "high tail," meaning that it converges to zero as the variate $x \to +\infty$ more slowly than an exponential distribution such as the Gaussian or Laplace distribution. The lognormal distribution is often used to represent the distribution of income or wealth in populations, for example.
5. A chi-squared distribution with two degrees of freedom, which has mean equal to 2 and variance equal to 4. This also has a "thick tail," but not a "high tail," meaning that it also converges slowly to zero as its argument approaches infinity—but not more slowly than exponential distributions. The chi-squared distribution is used in statistical hypothesis testing and for monitoring the behavior of Kalman filters.

The values of the mean-squared estimation error as a function of the value of the estimate are plotted in Figure 3.2, for all five of the distributions. These do not show

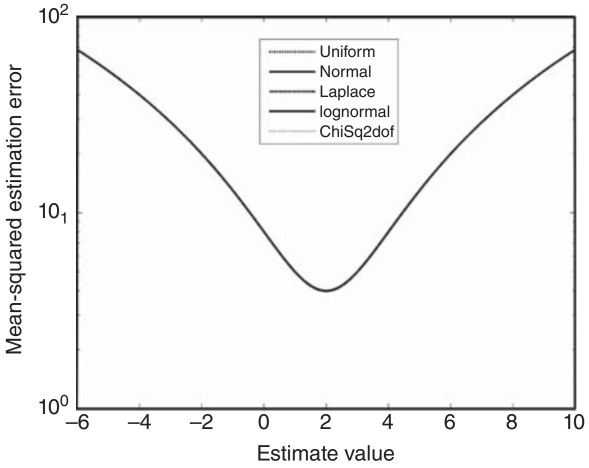

Figure 3.2 Mean-squared estimation error versus estimate value.

much difference between distributions, but the program `meansqesterr.m` (among the m-files on the Wiley web site) generated using MATLAB software also plots the differences—just to show that they are not exactly the same.

The mean-squared-error plots in Figure 3.2 all have the minimum mean-squared-error value of 4—the second central moment of all the distributions—at the estimated value of $\hat{x} = 2$—the mean of all the distributions.

The point is that the LMSE of the variate in all cases is the mean of the distribution, and the mean-squared estimation error is equal to the variance of the distribution—independent of the shape of the distribution.

3.4.4 Means and Covariances: Moments to Remember

Gamblers and actuaries are rightfully concerned about probability distributions, but distributions are not essential for least-mean-squares estimation. That is fortunate, because it is usually much easier and more efficient to estimate just the mean and covariance of a variate x.

3.4.4.1 Recursive Estimates of Means and Covariances The means and covariances of given data sequences $\{x[k] \mid k = 0, 1, 2, \ldots\}$ can be estimated recursively by the formulas

$$\hat{\mu}[k+1] = \hat{\mu}[k] + \frac{1}{k+1}[x[k+1] - \hat{\mu}_k] \tag{3.67}$$

$$\widehat{P}_{xx}[k+1] = \left(1 - \frac{1}{k}\right)\widehat{P}_{xx}[k] + (1+k)(\hat{\mu}[k+1] - \hat{\mu}[k])(\hat{\mu}[k+1] - \hat{\mu}[k])^{\mathrm{T}}$$

$$k = 0, 1, 2, 3, \ldots, \tag{3.68}$$

where $\hat{\mu}[k]$ is the kth recursive estimate of the mean and $\widehat{P}_{xx}[k]$ is the kth recursive estimate of the covariance, using the samples up to the kth element. To start the algorithm off, the value of $\hat{\mu}[0]$ can be initialized at $x[0]$, and the value of $\widehat{P}_{xx}[0]$ can be initialized as the zero matrix.

3.4.4.2 Means and Variances in Least-squares Sensor Calibration Least-squares estimation is used for removing known trends from sensor data, but the same solution can also provide an estimate of the variance of measurement or sensor error. This is also a by-product of the procedure for calibrating a sensor by observing its outputs while controlling its inputs. Such variances are a direct by-product from solving the associated least-squares problem using augmented upper-triangular[7] Cholesky's decomposition. In this approach, the linear system $Ax = b$ is solved by upper-triangular Choleskys decomposition of the augmented matrix symmetric product

$$[A \mid b]^{\mathrm{T}}[A \mid b] = \left[\frac{A^{\mathrm{T}}A \mid A^{\mathrm{T}}b}{\left(A^{\mathrm{T}}b\right)^{\mathrm{T}} \mid b^{\mathrm{T}}b}\right], \tag{3.69}$$

the result of which is an augmented upper-triangular *Choleskys factor* matrix $\mathcal{U}$ with partitioning

$$\mathcal{U} = \left[\frac{U \mid y}{0 \mid \varepsilon}\right], \tag{3.70}$$

where U is upper triangular and the symmetric product

$$\left[\frac{A^{\mathrm{T}}A \mid A^{\mathrm{T}}b}{\left(A^{\mathrm{T}}b\right)^{\mathrm{T}} \mid b^{\mathrm{T}}b}\right] = \mathcal{U}^{\mathrm{T}}\mathcal{U} \tag{3.71}$$

$$= \left[\frac{U^{\mathrm{T}} \mid 0}{0 \mid \varepsilon}\right]^{\mathrm{T}} \left[\frac{U \mid y}{0 \mid \varepsilon}\right] \tag{3.72}$$

$$= \left[\frac{U^{\mathrm{T}}U \mid U^{\mathrm{T}}y}{\left(U^{\mathrm{T}}y\right)^{\mathrm{T}} \mid |y|^2 + \varepsilon^2}\right], \tag{3.73}$$

the last line of which is a matrix equation which can be solved for the least-squares solution $\hat{x}$ as the solution of the Cholesky form of the *normal equation* for the least-squares problem:

$$U\hat{x} = y. \tag{3.74}$$

This avoids the problem of inverting a Gramian matrix, but adds the computational cost of the Cholesky decomposition. The Cholesky form of the normal equation can

[7] Most treatments of Cholesky's decomposition assume a lower-triangular result. However, the upper-triangular equivalent is just as efficient, accurate, and useful. It only means that the resulting back-substitution algorithm for inverting the result will start with the last row and progress backward.

be solved without matrix inversion—by using back substitution, instead—because U is triangular.

Furthermore, the yet-unused result $\varepsilon = |Ax - b|$, the root-sum-squared estimation error. In this case, the unbiased estimate for the standard deviation of the error in the data b would be $\sigma = \varepsilon / \sqrt{n-1}$, where n is the dimension of b. The mean error in the data is called *sensor bias.* It is usually part of the estimated vector $\hat{x}$.

This approach applies to trend removal by fitting input–output pairs of data with a known parametric functional form linear in the unknown parameters, such as fitting b to polynomials in a known input quantity χ. In this case, the elements of A are powers of the input variable χ corresponding to the output values b and the components of the vector x are the unknown coefficients of the polynomial.

3.4.5 Alternative Measures of Miss Distance

A *metric* is a measure of distance between two points in a topological space, which includes n-dimensional real spaces.

The Kalman filter uses the so-called "Euclidean metric," which is the square root of the sum of the squares of the components of the actual impact point relative to the intended target. It is based on the *Euclidean norm,* which is the square root of the sum of the squares of a vector. The Euclidean norm is also called the H_2 *norm*, a reference to the so-called "Hölder p-norm" for $p = 2$. In essence, this sort of metric is a norm applied to the difference between two points.

However, there are alternative metrics for $\Re^n$. One of these is the so-called "taxicab metric," based on the "sup-norm" or "H_∞ norm," defined as the maximum absolute component of a vector. The H_∞ norm leads to an alternative development of optimal estimation and control, also called *minimax* estimation and control. For the example of the archer, this would only make the bullseyes on the target be squares, not circles. For side-by-side development and comparison of minimax and Kalman filtering, see Reference 8.

3.5 TRANSFORMATIONS OF VARIATES

The critical question is: If a dependent variate Y is defined in terms of an existing variate X by a function f as

$$Y = f(X), \tag{3.75}$$

where the moments of X are known, what are the mean and covariance of Y?

This matters in Kalman filtering, where the variables of interest are the first and second moments of Y. The Kalman filter depends on f being linear, in which case the answer turns out to be straightforward and simple. The formulas for this case are derived in this section, along with formulas for what happens when f is not linear, but sufficiently smooth.

3.5.1 Linear Transformations

We will now put material from Section 3.3.2 in the context of general transformations of variates—without regard for the underlying distributions beyond their means and covariances.

If $x \in \mathfrak{R}^n$ is an n-dimensional real vector variate with probability density function $p(x)$ and A is a matrix conformable[8] for multiplication by x, the dependent variate

$$y = Ax \tag{3.76}$$

will have mean

$$\mu_y = \underset{x}{\mathrm{E}}\langle y\rangle = \int Ax\, p(x)\, dx = A\int x\, p(x)\, dx = A\mu_x \tag{3.77}$$

and covariance

$$P_{yy} = \underset{x}{\mathrm{E}}\langle (y-\mu_y)(y-\mu_y)^{\mathrm{T}}\rangle \tag{3.78}$$

$$= \underset{x}{\mathrm{E}}\langle (Ax - A\mu_x)(Ax - A\mu_x)^{\mathrm{T}}\rangle \tag{3.79}$$

$$= \underset{x}{\mathrm{E}}\langle A(x-\mu_x)(x-\mu_x)^{\mathrm{T}}A^{\mathrm{T}}\rangle \tag{3.80}$$

$$= A\,\underset{x}{\mathrm{E}}\langle (x-\mu_x)(x-\mu_x)^{\mathrm{T}}\rangle A^{\mathrm{T}} \tag{3.81}$$

$$= AP_{xx}A^{\mathrm{T}}. \tag{3.82}$$

These results do not depend on the details of the probability distributions but only on their first moments and second central moments. That is, any linear transformation of the variates transforms the mean and covariance matrix of the distribution that is exactly the same, independent of the underlying probability distribution.

3.5.1.1 Linear Combinations of Vector Variates Given the definition and properties of cross-covariance and statistical independence, one can uncover some useful properties of linear combinations of vector variates, where the coefficients of the linear combination are now matrices.

If X_a and X_b are vector variates with some as-yet unspecified joint probability distribution, then the composite sample vectors

$$x \overset{\text{def}}{=} \begin{bmatrix} x_a & \in & X_a \\ x_b & \in & X_b \end{bmatrix} \tag{3.83}$$

[8] A notation defined in Appendix B on the companion Wiley web site, and meaning the data structure in question has the correct dimensions for the way it is being used.

will have covariance matrix

$$P_{xx} = \begin{bmatrix} P_{x_a x_a} & P_{x_a x_b} \\ P_{x_b x_a} & P_{x_b x_b} \end{bmatrix}. \tag{3.84}$$

Any linear combination

$$y = A\ x_a + B\ x_b \tag{3.85}$$

$$= \begin{bmatrix} A & B \end{bmatrix} x \tag{3.86}$$

will then have mean

$$\mu_y = A\ \mu_{x_a} + B\ \mu_{x_b} \tag{3.87}$$

and covariance

$$P_{yy} = \begin{bmatrix} A & B \end{bmatrix} \begin{bmatrix} P_{x_a x_a} & P_{x_a x_b} \\ P_{x_b x_a} & P_{x_b x_b} \end{bmatrix} \begin{bmatrix} A & B \end{bmatrix}^{\mathrm{T}} \tag{3.88}$$

$$= AP_{x_a x_a} A^{\mathrm{T}} + AP_{x_a x_b} B^{\mathrm{T}} + BP_{x_b x_a} A^{\mathrm{T}} + BP_{x_b x_b} B^{\mathrm{T}}. \tag{3.89}$$

In the case that X_a and X_b are *statistically independent,*

$$P_{yy} = AP_{x_a x_a} A^{\mathrm{T}} + BP_{x_b x_b} B^{\mathrm{T}}. \tag{3.90}$$

3.5.1.2 Affine Transformations An *affine transformation* is essentially the first two terms (zeroth order and first order) of a series expansion of a function. It is equivalent to a linear transformation plus an offset, such as

$$y = Ax + b, \tag{3.91}$$

where b is a conformable constant vector.

For any variate x with mean μ_x, the mean of y after such an affine transformation will be

$$\mu_y \overset{\text{def}}{=} \underset{x}{\mathrm{E}}\langle y \rangle \tag{3.92}$$

$$= \underset{x}{\mathrm{E}}\langle Ax \rangle + \underset{x}{\mathrm{E}}\langle b \rangle \tag{3.93}$$

$$= A\,\underset{x}{\mathrm{E}}\langle x \rangle + b \tag{3.94}$$

$$= A\mu_x + b. \tag{3.95}$$

Similarly, its covariance

$$P_{yy} \overset{\text{def}}{=} \underset{x}{\mathrm{E}}\langle (y - \mu_y)(y - \mu_y)^{\mathrm{T}} \rangle \tag{3.96}$$

$$= \mathop{\mathrm{E}}_{x}\langle (Ax + b - A\mu_x - b)(Ax + b - A\mu_x -)^{\mathrm{T}}\rangle \tag{3.97}$$

$$= \mathop{\mathrm{E}}_{x}\langle A(x - \mu_x)(x - \mu_x)^{\mathrm{T}}A^{\mathrm{T}}\rangle \tag{3.98}$$

$$= AP_{xx}A^{\mathrm{T}}. \tag{3.99}$$

In summary, then, affine transformations add a constant bias to the mean, but otherwise behave the same as linear transformations.

3.5.1.3 Independent Random Offsets Linearity also extends to linear combinations of independent random vector variates.

Let the dependent variate

$$y = Ax + Bw, \tag{3.100}$$

where x is a vector variate with mean μ_x and w is a statistically independent variate with mean μ_w. Then the mean of y

$$\mu_y = \mathop{\mathrm{E}}_{x,w}\langle y\rangle \tag{3.101}$$

$$= \mathop{\mathrm{E}}_{x}\langle Ax\rangle + \mathrm{E}_w\langle Bw\rangle \tag{3.102}$$

$$= A\mathop{\mathrm{E}}_{x}\langle x\rangle + B\mathrm{E}_w\langle w\rangle \tag{3.103}$$

$$= A\mu_x + B\mu_w \tag{3.104}$$

and the covariance

$$P_{yy} = \mathop{\mathrm{E}}_{x,w}\langle (Ax + Bw - A\mu_x - B\mu_w)(Ax + Bw - A\mu_x - B\mu_w)^{\mathrm{T}}\rangle \tag{3.105}$$

$$= \mathop{\mathrm{E}}_{x}\langle A(x - \mu_x)(x - \mu_x)^{\mathrm{T}}A^{\mathrm{T}}\rangle + \mathrm{E}_w\langle B(w - \mu_w)(w - \mu_w)^{\mathrm{T}}B^{\mathrm{T}}\rangle \tag{3.106}$$

$$= AP_{xx}A^{\mathrm{T}} + BP_{ww}B^{\mathrm{T}}. \tag{3.107}$$

This shows the result of adding independent random noise to a vector variate.

3.5.2 Transformations by Analytic Functions

The formulas for linear transformations of vector variates are simple and straightforward. It is not so simple when the transformations are nonlinear.

We consider here the case in which the transformations can be represented in terms of a power series of the variate. This includes the affine case when the coefficients are zero beyond the first-order term in the series, but it also shows how things unravel beyond that.

3.5.2.1 The Scalar Case Scalar analytic functions are defined by power series,

$$f(x) = \sum_{k=0}^{\infty} a_k x^k. \tag{3.108}$$

Transformation of the Mean The expected value

$$\underset{x}{\mathrm{E}}\langle f(x)\rangle = \underset{x}{\mathrm{E}}\left\langle \sum_{k=0}^{\infty} a_k x^k \right\rangle \tag{3.109}$$

$$= \sum_{k=0}^{\infty} a_k \underset{x}{\mathrm{E}}\langle x^k\rangle \tag{3.110}$$

$$= \sum_{k=0}^{\infty} a_k \, {}^{[k]}\mu_x \tag{3.111}$$

$${}^{[k]}\mu_x \overset{\text{def}}{=} \underset{x}{\mathrm{E}}\langle x^k\rangle, \tag{3.112}$$

the kth raw moment of the underlying probability distribution.

Equation 3.111 then defines the LMSE of the value of an analytic function in terms of the coefficients of the power series expansion of the function and the moments of the probability distribution.

This shows that the mean of a variate after an analytic transformation depends on the original moments of all orders, weighted by the coefficients of the power series expansion.

Transformation of Covariance The other moment of interest is the value of the mean-squared estimation error,

$$\underset{x}{\mathrm{E}}\langle (f(x) - \underset{x}{\mathrm{E}}\langle f(x)\rangle)^2\rangle = \sum_{k=0}^{\infty} \left[\sum_{j=0}^{k} a_j^{\star} a_{k-j}^{\star} \right] {}^{[k]}\mu_x \tag{3.113}$$

$$a_0^{\star} = a_0 - \sum_{k=0}^{\infty} a_k \, {}^{[k]}\mu_x \tag{3.114}$$

$$a_j^{\star} = a_j, \ j > 0, \tag{3.115}$$

showing that, in the case of an analytic function, the mean-squared estimation error also involves all orders of moments of the initial probability distribution.

Example 3.8 (Quadratic Transformation of a Scalar Gaussian Variate) Let $x \in X = \mathcal{N}(\mu_x, P_{xx})$ for scalar μ_x and P_{xx}, and let the dependent variate Y be realized as

$$y = y_0 + y_1 x + y_2 x^2,$$

where the scalar constants y_0, y_1, and y_2 are known.

Then the mean of Y

$$\begin{aligned}
\mu_y &= \underset{x}{\mathrm{E}}\langle y_0 + y_1 x + y_2 x^2\rangle \\
&= y_0 + y_1\mu_x + y_2 \underset{x}{\mathrm{E}}\langle[(x-\mu_x)+\mu_x]^2\rangle \\
&= y_0 + y_1\mu_x + y_2[\underset{x}{\mathrm{E}}\langle(x-\mu_x)^2\rangle + 2\,\underset{x}{\mathrm{E}}\langle(x-\mu_x)\mu_x\rangle + \mu_x^2] \\
&= y_0 + y_1\mu_x + y_2[P_{xx} + \mu_x^2]
\end{aligned}$$

now involves the covariance of X.

Worse yet, the covariance of y

$$\begin{aligned}
P_{yy} &= \underset{x}{\mathrm{E}}\langle(y_0 + y_1 x + y_2 x^2 - \mu_y)^2\rangle \\
&= \underset{x}{\mathrm{E}}\langle(y_0 + y_1 x + y_2 x^2 - y_0 - y_1\mu_x - y_2 P_{xx} - y_2\mu_x^2)^2\rangle \\
&= \underset{x}{\mathrm{E}}\langle[y_1(x-\mu_x) + y_2(x^2 - P_{xx} - \mu_x^2)]^2\rangle \\
&= y_1^2\,\underset{x}{\mathrm{E}}\langle(x-\mu_x)^2\rangle \\
&\quad +2y_1 y_2\,\underset{x}{\mathrm{E}}\langle(x-\mu_x)(x^2 - P_{xx} - \mu_x^2)\rangle \\
&\quad +y_2^2\,\underset{x}{\mathrm{E}}\langle(x^2 - P_{xx} - \mu_x^2)^2\rangle \\
&= y_1^2 P_{xx} + 2y_1 y_2\,\underset{x}{\mathrm{E}}\langle(x-\mu_x)[(x-\mu_x)(x+\mu_x) - P_{xx}]\rangle \\
&\quad +y_2^2 E_x\langle[(x-\mu_x)(x+\mu_x) - P_{xx}]^2\rangle
\end{aligned}$$

now involves moments up to fourth order. One can complete this formula by plugging in all the moments of the original Gaussian distribution, which only depend on μ_x and P_{xx}.

The requisite higher order moments from Table 3.1 could be plugged into the last formula to express everything in terms of the original Gaussian first- and second-order central moments μ and σ^2. However, the resulting distribution will no longer be Gaussian, and the resulting formula using Gaussian moments would not work the next time a nonlinear transformation is applied.

3.5.2.2 The Vector Case An affine transformation includes just the first two terms of a vector-valued power series, the first (zeroth-order) term of which is a constant offset vector ${}^{[0]}a$ and the second (first-order) term of which is a linear transformation, characterized by a two-dimensional data array (matrix) $^{[1]}A$:

$$f(x) = {}^{[0]}a + {}^{[1]}Ax + \cdots \tag{3.116}$$

$$= \begin{bmatrix} {}^{[0]}a_1 \\ {}^{[0]}a_2 \\ {}^{[0]}a_3 \\ \vdots \\ {}^{[0]}a_n \end{bmatrix} + F_1 x + \cdots \ , \tag{3.117}$$

with the contribution to the ith component of $f(x)$ being

$$f_i(x) = \ {}^{[0]}a_i + \sum_{j=1}^{n} {}^{[1]}A_{i,j} x_j + \cdots \ . \tag{3.118}$$

The next (second-order) term of that series uses a three-dimensional data array $^{[2]}A$, with its contribution to the ith column of $f(x)$ being

$$f_i(x) = F_{0,i} + \sum_{j=1}^{n} F_{1,i,j} x_j + \sum_j \sum_k F_{2,i,j,k} x_j x_k + \cdots \ . \tag{3.119}$$

The next higher order (third-order) term will have the form

$$f_i(x) = F_{0,i} + \sum_{j=1}^{n} F_{1,i,j} x + j + \sum_j \sum_k F_{2,i,j,k} x_j x_k + \sum_j \sum_k \sum_\ell F_{3,i,j,k,\ell} x_j x_k x_\ell + \cdots \ , \tag{3.120}$$

and so forth for all higher order terms. Each successive term uses a data array of one dimension greater than the previous term, and the implementation has one more summation. The data requirements just to hold the multidimensional coefficient arrays will then be n^k for the kth term, or

$$\sum_{k=1}^{N} n^k = \begin{cases} N, & n = 1 \\ \dfrac{n\left(n^N - 1\right)}{(n-1)}, & n > 1 \end{cases} \tag{3.121}$$

for the first N terms.

The kth term then requires $\mathcal{O}(k\ n^k)$ arithmetic operations for evaluation (counting the multiplications of powers of the variates), and an expansion to Nth order would require something in the order of

$$\sum_{k=1}^{N} k\ n^k = \begin{cases} N\,(N+1)\,/2, & n = 1 \\ -\dfrac{Nn^{N+1}}{1-n} + \dfrac{n(1-n^N)}{(1-n)^2}, & n > 1, \end{cases} \tag{3.122}$$

arithmetic operations.

Impact on Means and Covariances The down side of all this is that the higher order terms in the transformation require more data arrays in the model, and more central moments of the underlying probability distribution beyond the covariance—just for computing the covariances used in Kalman filtering.

For example, with just one additional (second-order) term in the expansion, the term in the ith row and jth column of the covariance matrix of $f(x)$

$$\begin{aligned} P_{ff,i,j} &= \mathop{\mathrm{E}}_{x}\langle f_i(x) f_j(x)\rangle \qquad (3.123)\\ &= \mathop{\mathrm{E}}_{x}\left\langle \left[F_{0,i} + \sum_{k=1}^{n} F_{1,i,k}x_k + \sum_{k}\sum_{\ell} F_{2,i,k,\ell}x_k x_\ell\right]\right.\\ &\quad \times \left.\left[F_{0,j} + \sum_{k=1}^{n} F_{1,j,k}x + j + \sum_{k}\sum_{\ell} F_{2,j,k,\ell}x_k x_\ell\right]\right\rangle, \qquad (3.124) \end{aligned}$$

which now includes taking the expected value of terms up to fourth order. That is, computing the covariance matrix of a second-order transformation of a variate requires knowing the central moments of its distribution up to fourth order.

In general, computing the covariance matrix of an Nth-order transformation of a variate requires knowing the central moments up to $(2N)$th order. If the resulting distribution were Gaussian, all central moments can be derived from just the mean and covariance. However, nonlinear transformations of Gaussian distributions are no longer Gaussian.

That can be a bit discouraging, when all we want are means and covariance matrices.

3.5.3 Transformation of Probability Density Functions

Kalman filtering is not much concerned with probability density functions, except as an intermediary step for understanding the properties of the expectancy operator $\mathrm{E}x\langle\cdot\rangle$. However, it is sometimes useful to understand how transformations of the variate transform the probability density functions.

3.5.3.1 Linear Transformations Linear transformations of variates defined on the real line $\mathfrak{R}$ are defined by $Y = aX$, where $a \neq 0$. If the variate X has probability density function $p_x(\cdot)$, then the probability density function of Y will be defined by

$$p_y(y) = |a|^{-1} p_x(a^{-1}y). \qquad (3.125)$$

3.5.3.2 Nonlinear Transformations If the scalar variate Y is defined by $y = f(x)$, $x \in X$ with probability density $p_x(\cdot)$, the derivation is not so simple as in the linear case. However, if $f(\cdot)$ and its inverse function $f^{-1}(\cdot)$ are differentiable everywhere, then the probability density function of Y is given by

$$p_y(y) = \left|\frac{\partial f^{-1}(y)}{\partial y}\right| p_x(f^{-1}(y)) \tag{3.126}$$

$$= \frac{p_x(f^{-1}(y))}{\left|\left.\frac{\partial f(x)}{\partial x}\right|_{x\,=\,f^{-1}(y)}\right|}, \tag{3.127}$$

where the absolute value $|\cdot|$ is needed to keep the probability density function $p_y \geq 0$, and this only works where the derivative of $f(\cdot) \neq 0$.

Note that this is equivalent to Equation 3.125 in the case that $f(x) = ax$.

Example 3.9 (Arctangent Transformation of Univariate Gaussian Distribution) Let

$$y = f(x)$$

$$= \arctan\,(a\,x)$$

$$f^{-1}(y) = \frac{1}{a}\tan\,(y),$$

where $x \in X$, a zero-mean unit normal distribution with probability density function

$$p_x(x) = \frac{\exp\,(-x^2/2)}{\sqrt{2\pi}}.$$

Note that both $f(\cdot)$ and $f^{-1}(\cdot)$ are differentiable, and the derivative of $f^{-1}(y) > 0$ and finite for $-\frac{\pi}{2} < y < \frac{\pi}{2}$. Applying Equation 3.126,

$$p_y(y) = \left|\frac{1}{a}\frac{\partial\tan\,(y)}{\partial y}\right| p_x\left(\frac{1}{a}\tan\,(y)\right)$$

$$= \frac{1}{a\sqrt{2\pi}}\{1 + [\tan(y)]^2\}\exp\left(-\frac{1}{2}\left[\frac{\tan\,(y)}{a}\right]^2\right), \quad -\frac{\pi}{2} < y < \frac{\pi}{2}.$$

It has the shapes shown in Figure 3.3 for various values of the positive parameter a. These might resemble a Gaussian distribution when $a \ll 1$, but are decidedly non-Gaussian and bimodal for $a \gtrapprox 1$.

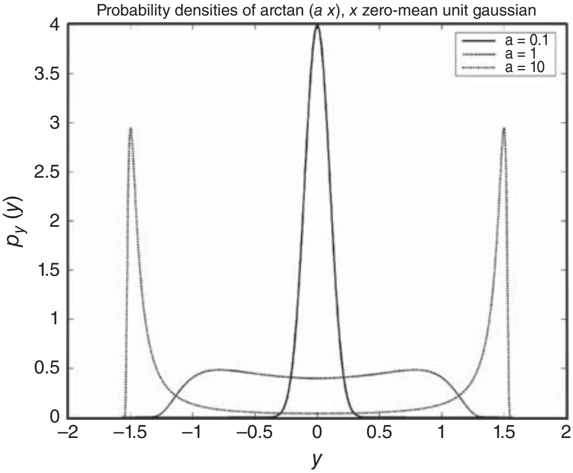

Figure 3.3 Probability densities of $y = \arctan(ax)$, X zero-mean unit-Gaussian.

3.6 THE MATRIX TRACE IN STATISTICS

The *trace* of a square matrix A is defined as the sum of its diagonal elements:

$$\text{tr}\,[A] \overset{\text{def}}{=} \sum_{i=1}^{n} a_{ii}. \tag{3.128}$$

It has some properties found to be useful in statistics—and in Kalman filtering.

3.6.1 Connecting Covariances and Mean-Squared Magnitudes

If P is the covariance (second central moment) of n-vector variate X, then its trace

$$\text{tr}\,[P] = \sum_{i=1}^{n} p_{ii} \tag{3.129}$$

$$= \sum_{i=1}^{n} \mathop{\text{E}}_{x} \langle (x_i - \mu_{x_i})^2 \rangle \tag{3.130}$$

$$= \mathop{\text{E}}_{x} \langle |x - \mu_x|^2 \rangle, \tag{3.131}$$

the mean-squared magnitude of $x - \mu_x$.

3.6.2 A Linear Functional

It we consider an $n \times n$ matrix to be a real function defined on the pairs of integers $\{(i, j) | 1 \le i \le n,\ 1 \le j \le n\}$, then the trace becomes a linear functional:

$$\begin{aligned} \text{tr}\,[A + B] &= \sum_{i=1}^{n} (a_{ii} + b_{ii}) && (3.132)\\ &= \sum_{i=1}^{n} a_{ii} + \sum_{i=1}^{n} b_{ii} && (3.133)\\ &= \text{tr}\,[A] + \text{tr}\,[B] && (3.134)\\ \text{tr}\,[cA] &= \sum_{i=1}^{n} c\, a_{ii} && (3.135)\\ &= c \sum_{i=1}^{n} a_{ii} && (3.136)\\ &= c\ \text{tr}\,[A]. && (3.137) \end{aligned}$$

3.6.3 Matrix Product Commutation Under the Trace

If the matrix A is $n \times n$ and the matrix B is $m \times n$, then their products AB and BA are both square matrices. (AB is $n \times n$ and BA is $m \times m$.) Therefore, either product has a matrix trace.

What is even more profound is that their traces are equal:

$$\begin{aligned} \text{tr}\,[AB] &= \sum_{i=1}^{n} \underbrace{\sum_{j=1}^{m} a_{ij} b_{ji}}_{\{AB\}_{ii}} && (3.138)\\ &= \sum_{j=1}^{m} \underbrace{\sum_{i=1}^{n} b_{ji} a_{ij}}_{\{BA\}_{jj}} && (3.139)\\ &= \text{tr}\,[BA]. && (3.140) \end{aligned}$$

That is, the trace of AB equals the trace of BA.

In other words, cyclical permutation of matrix factors under the trace operator can be allowed:

$$\begin{aligned} \text{tr}\,[A_1 \times A_2 \times A_3 \times \cdots \times A_N] &= \text{tr}\,[A_2 \times A_3 \times \cdots \times A_N \times A_1] && (3.141)\\ &= \text{tr}\,[A_N \times A_1 \times A_2 \times A_3 \times \cdots \times A_{N-1}], && (3.142) \end{aligned}$$

and these formulas can be applied repeatedly to obtain any cyclical permutation of the factors.

As a consequence, if $\nu \in \mathcal{N}(0,\ P_{\nu\nu})$, then the expected value

$$
\begin{aligned}
\mathrm{E}_{\nu}\langle \nu^{\mathrm{T}} P_{\nu\nu}^{-1} \nu \rangle &= \mathrm{E}_{\nu}\langle \mathrm{tr}\ [\nu^{\mathrm{T}} P_{\nu\nu}^{-1} \nu] \rangle && (3.143)\\
&= \mathrm{E}_{\nu}\langle \mathrm{tr}\ [P_{\nu\nu}^{-1} \nu\nu^{\mathrm{T}}] \rangle \text{(trace property)} && (3.144)\\
&= \mathrm{tr}\ [P_{\nu\nu}^{-1} \mathrm{E}_{\nu}\langle \nu\nu^{\mathrm{T}} \rangle] && (3.145)\\
&= \mathrm{tr}\ [P_{\nu\nu}^{-1} P_{\nu\nu}] && (3.146)\\
&= \mathrm{tr}\ [I_{\ell}] && (3.147)\\
&= \ell, && (3.148)
\end{aligned}
$$

the dimension of the variate ν. If properly modeled and the noise is Gaussian, the sequence of values $\{\nu_k^{\mathrm{T}} P_{\nu\nu\ k}^{-1} \nu_k\}$ should have a unit chi-squared distribution with ℓ degrees of freedom. The mean of this distribution is ℓ—as shown above—and the variance is 2ℓ.

3.6.4 Chi-Squared Test

The "chi-squared" (χ^2) test [8] was designed to test whether a given sequence $\{\nu_k\}$ of zero-mean white Gaussian noise is from a Gaussian distribution with a given covariance $P_{\nu\nu}$. If so, the scalar variates

$$\xi(\nu_k) \overset{\mathrm{def}}{=} \nu_k^{\mathrm{T}} P_{\nu\nu}^{-1} \nu_k \tag{3.149}$$

should have a unit chi-squared distribution with ℓ degrees of freedom, where ℓ is the dimension of the ν_k.

The probability density function for a unit chi-squared (χ^2) variate $\nu > 0$ with ℓ degrees of freedom is

$$p_{\chi^2}(\nu,\ \ell) = \frac{\nu^{\ell/2-1} e^{-\nu/2}}{2^{\ell/2}\Gamma(\ell/2)}, \tag{3.150}$$

which can be calculated using the MATLAB function `chi2pdf` in the Statistics Toolbox, or using the GNU[9] version with the same name. This MATLAB function computes the sequence of probability densities for a given sequence of samples. The MATLAB function `prod` converts the array of probability densities into a joint probability density for the entire sequence.

Given N such samples of χ^2 variates, this procedure yields the relative probability density function for each of the samples. The statistical decision in this case will be that data sequence with the highest joint probability density.

[9] Distributed by the Free Software Foundation, which claims that "GNU" stands for "GNU's Not Unix."

3.6.5 Schweppe Likelihood Ratio Detection

The chi-square test can also be used when there are more than two possibilities for the distribution behind the observed samples.

Schweppe [9] applied this to the case in which there are two competing hypotheses:

1. A measured signal comes only from a noise source with a known stochastic dynamic model.
2. The measured signal is the sum of the above noise source, plus a signal with known distinct stochastic structure.

In this case, the solution has come to be called *Schweppe Gaussian likelihood ratio detection*. It answers the question of which of two different Gaussian linear stochastic system models ("noise-only" model or a "signal + noise" model) is more likely to be the best model for an observed sequences of sampled values.

It is a form of likelihood ratio test using the linear stochastic models of Kalman filtering. The approach subsumes Kalman filtering, but the test itself only involves testing of white noise processes with one of the two given stochastic models: one with a "noise-only" model and another with a "signal + noise" model. If only noise is present, the sequence of differences between the filter-predicted values and the measured values should be a zero-mean Gaussian white noise sequence

$$\{\nu_k^{[N]}|\ k = 1,\ 2,\ 3,\ \ldots\}$$

with known covariance $P_{\nu\nu}^{[N]}$ (computed by the Kalman filter).

If, on the other hand, the "signal + noise" model is the correct one, the corresponding sequence

$$\{\nu_k^{[S]}\ |\ k = 1,\ 2,\ 3,\ \ldots\}$$

should be a zero-mean Gaussian white noise sequence with known covariance $P_{\nu\nu}^{[S]}$—also computed by its Kalman filter.

Signal detection is then a decision process comparing the relative probabilities of the two sequences of differences between the predicted and observed measurements from two Kalman filters. In either case, if properly modeled, the sequence of outputs should be a zero-mean white (uncorrelated) Gaussian noise process with known covariance. The decision process compares the respective Gaussian probability densities with the assumed covariance matrices for each model. The Gaussian probability density function $p(\nu)$ for $\nu \in \mathcal{N}(0,\ P_{\nu\nu})$ in either case is given by Equation 3.3, and the ratio of the probability densities

$$\frac{p(\nu_k^{[N]})}{p(\nu_k^{[S]})} = \frac{\sqrt{\det P_{\nu\nu}^{[S]}}\ \exp\ \left(-\frac{1}{2}\nu_k^{[N]\mathrm{T}}\left(P^{[N]}\right)^{-1}\nu_k^{[N]}\right)}{\sqrt{\det P_{\nu\nu}^{[S]}}\ \exp\ \left(-\frac{1}{2}\psi_i^{\mathrm{T}}\left(P^{[S]}\right)^{-1}\psi_i\right)} \tag{3.151}$$

$$\log \left[\frac{p\left(\nu_i\right)}{p(\psi_i)}\right] = -\frac{1}{2}\nu_k^{[N]\mathrm{T}}(P^{[N]})^{-1}\nu_k^{[N]} + \frac{1}{2}\nu_k^{[S]\mathrm{T}}(P^{[S]})^{-1}\nu_k^{[S]}$$
$$+\frac{1}{2}[\log\det P_{\nu\nu}^{[S]} - \log\det P_{\nu\nu}^{[S]}], \tag{3.152}$$

which can be used to guess which sequence is more likely to have P as its covariance.

If the sequence in question is truly white, the joint probability density of a partial sequence of outputs is the product of the individual probability densities, and logarithm of that joint probability density is the sum of the logarithms of the individual probability densities.

3.6.6 Multihypothesis Detection

This approach can be extended to selecting which of $N > 2$ such statistical sequences is most likely, something which has been used in extended Kalman filtering to overcome linearization errors in the initial (detection) phase of detection and tracking (see Chapter 8).

3.7 SUMMARY

1. Probability theory began with questions about the odds of discrete outcomes (especially in gambling), but progressed to problems with inputs and outputs characterized by real numbers.
2. Probability distributions defined on real domains are defined in terms of probability *measures*—a wholly different breed of cat.
3. Kalman filtering is based on probability distributions defined on n-dimensional real vector spaces. The statistical characteristics of these probabilities defined in this chapter would apply as well to applications defined on closed n-dimensional topological manifolds without boundaries, such as angles defined on the circle or attitudes defined on the three-dimensional surface of the unit sphere in four-dimensional quaternion space. Statistical properties of probability distributions are not an issue with applications defined on $\mathfrak{R}^n$ or on n-dimensional manifolds other than $\mathfrak{R}^n$.
4. The critical variables used in Kalman filtering can be identified with the first two *moments* of the underlying probability distributions.
5. The first moment is called the *mean* of the distribution. It is an n-dimensional vector.
6. The mean is also the *estimate* of the variate with the *least-mean-squared estimation error*. This result does not depend on what probability distribution is used, so long as it has the required first and second moments.
7. The second moment of the deviation from the mean is called the *covariance* of the distribution, also called the *second central moment* of the distribution. It is an $n \times n$ symmetric positive-semi-definite matrix.

8. Subvectors of a vector variate are *statistically independent* if and only if their cross-covariances are zero.
9. The covariance matrix characterizes the *minimum-mean-squared estimation error.*
10. Linear transformations $y = Ax$ of an n-dimensional vector variate x with mean μ_x results in a y probability distribution in which the transformed mean $\mu_y = A\mu_x$—a result which depends only on μ_x and the transforming matrix A and not on any other attributes of the distribution.
11. The covariance matrix of the resulting y-distribution will be $P_{yy} = AP_{xx}A^{\mathrm{T}}$, where P_{xx} is the covariance of the x-distribution.
12. Any linear combination $y = Ax_a + Bx_b$ of vector variates X_a and X_b will have mean $\mu_y = A\mu_{x_a} + B\mu_{x_b}$ and covariance

$$P_{yy} = AP_{x_ax_a}A^{\mathrm{T}} + AP_{x_ax_b}B^{\mathrm{T}} + BP_{x_bx_a}A^{\mathrm{T}} + BP_{x_bx_b}B^{\mathrm{T}}.$$

13. Nonlinear transformations of probability distributions couple higher order moments into the first two moments.
14. The number of moments involved in evaluating the covariance of an Nth-order polynomial transformation of a variate grows as $2 \times N$.
15. The cumulative size of the data structures for the coefficients of a power series expansion of an n-vector to order N grows as

$$\frac{n(n^N - 1)}{(n-1)},$$

where n is the vector dimension and N is the highest power of the vector finite power series expansion. The number of arithmetic operations required to evaluate the expansion grows as

$$\frac{Nn^{N+1}}{n-1} + \frac{n(1-n^N)}{(n-1)^2}.$$

PROBLEMS

3.1 A deck of 52 playing cards is divided into four equal "suits" of 13 cards each, with each suit labeled with a heart (♡), diamond (◇), club (♣), or spade (♠). The 13 cards within each suit are labeled with an "A" (ace), 2, 3, 4, 5, 6, 7, 8, 9, 10, "J" (jack), "Q" (queen), or "K" (king). In a fair deal from a full deck, each card is equally likely to be drawn.

(a) In drawing a single card, what is the probability of drawing a spade?

(b) An ace?

(c) The ace of spades?

3.2 Let a deck of 52 cards be divided into four piles (labeled North, South, East, West). Find the probability that each pile contains exactly one ace.

3.3 Show that

$$\binom{n+1}{k+1} = \binom{n}{k+1} + \binom{n}{k},$$

for all $n > k$.

3.4 How many ways are there to divide a deck of 52 cards into four piles of 13 each?

3.5 If a hand of 13 cards are drawn from a deck of 52, what is the probability that exactly 3 cards are spades?

3.6 If the 52 cards are divided into four piles of 13 each, and if we are told that North has exactly three spades, find the probability that South has exactly three spades.

3.7 A hand of 13 cards is dealt from a well-randomized deck.

(a) What is the probability that the hand contains exactly seven hearts?

(b) During the deal, the face of one of the cards is inadvertently exposed and it is seen to be a heart. What is now the probability that the hand contains exactly seven hearts?

You may leave the above answers in terms of factorials.

3.8 The random variables $X_1, X_2, \ldots, X_n$ are independent with mean zero and the same variance σ_X^2. We define the new random variables $Y_1, Y_2, \ldots, Y_n$ by

$$Y_n = \sum_{j=1}^{n} X_j.$$

Find the correlation coefficient $\rho_{n-1,\,n}$ between Y_{n-1} and Y_n.

3.9 The random variables X and Y are independent and uniformly distributed between 0 and 1 (rectangular distribution). Find the probability density function of $Z = |X - Y|$.

3.10 Two random variables x and y have the density function

$$p_{xy}(x, y) = \begin{cases} C(y - x + 1), & 0 \le y \le x \le 1, \\ 0, & \text{elsewhere,} \end{cases}$$

where the constant $C < 0$ is chosen to normalize the distribution.

(a) Sketch the density function in the x, y plane.

(b) Determine the value of C for normalization.

(c) Obtain two marginal density functions.

(d) Obtain $\mathrm{E}\langle Y|x\rangle$.

(e) Discuss the nature and use of the relation $y = \mathrm{E}\langle Y|x\rangle$.

3.11 The random variable X has the probability density function

$$f_X(x) = \begin{cases} 2x, & 0 \le x \le 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the following:

(a) The cumulative function $F_X(x)$.

(b) The median.

(c) The mode.

(d) The mean, $\mathrm{E}\langle X\rangle$.

(e) The mean-square value $\mathrm{E}\langle X^2\rangle$.

(f) The variance $\sigma^2[X]$.

3.12 Can a probability distribution on a domain $\mathcal{D}$ with integrable functions $\mathcal{F}$ be defined in terms of a linear functional $\mathcal{P}$ such that

(a) $\mathcal{P}(f) \ge 0$ for all $f \in \mathcal{F}$ with nonnegative values, and

(b) $\mathcal{P}(1) = 1$, where the function "1" has value 1 everywhere in $\mathcal{D}$?

3.13 Derive and sketch the cumulative probability function for the Dirac δ distribution defined in Example 3.2.

3.14 For a univariate Gaussian probability density with mean μ and variance σ^2,

(a) What is the value of its *mode* (argument of probability density function at its maximum probability density)?

(b) What is the value of its *median* (argument of probability density function at which the cumulative probability equals $\frac{1}{2}$)?

(c) What is the LMSE estimate $\hat{x}_{\mathrm{LMSE}}$ of $x \in \mathcal{N}(\mu,\ \sigma^2)$?

3.15 Sketch the erf function, using an arrow to show where it has the value zero.

3.16 Write a formula for the cumulative probability function of the univariate Gaussian probability density with mean μ and variance σ^2, using the erf function.

3.17 The m-file `pYArctanaX.m` on the companion Wiley web site computes the function p_y given in Example 3.9. Write a script using MATLAB software calling `pYArctanaX(y,a)` to compute and plot the mean and variance of the resulting probability distribution as a function of the parameter a over the ranges $-\pi/2 \le y \le +\pi/2$ and $0.1 \le a \le 10$.

3.18 Use the m-file `pYArctanaX.m` with $a = 10$ (bimodal distribution) and MATLAB software to compute and plot the mean-squared estimation error as a function of the estimate $\hat{x}$ over the range $-\pi/2 \le \hat{x} \le +\pi/2$. Where does it achieve its minimum value?

REFERENCES

[1] P. Billingsley, *Probability and Measure, Anniversary Edition*, John Wiley & Sons, Inc., New York, 2012.

[2] C. M. Grinstead and J. L. Snell, *Introduction to Probability*, 2nd ed., American Mathematical Society, Providence, RI, 1997.

[3] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, New York, 2002.

[4] D. Shafer and V. Vovk, "The Sources of Kolmogorov's *Grundbegriffe*," *Statistical Science*, Vol. 21, No. 1, pp. 70–98, 2006.

[5] A. N. Kolmogorov, *Bundbegriffe der Wahrscheinlichkeitsrechnung*, Springer, Berlin, 1933.

[6] O. Hölder, "Ueber einen Mittelwertsatz," *Nachrichten von der Königl, Gesellschaft der Wissenschaften und der Georg-Augusts-Universität zu Göttingen*, No. 2, pp. 38–47, Band (Vol.) 1889.

[7] D. Simon, *Optimal State Estimation: Kalman, H_∞, and Nonlinear Applications*, John Wiley & Sons, Inc., Hoboken, NJ, 2006.

[8] P. E. Greenwood and S. N. Nikulin, *A Guide to Chi-Squared Testing*, John Wiley & Sons, Inc., New York, 1996.

[9] F. C. Schweppe, "Evaluation of likelihood functions for Gaussian signals," *IEEE Transactions on Information Theory*, Vol. IT-11, pp. 61–70, 1965.