Fundamentals of Sensor Fusion

Target tracking, navigation and SLAM

Edmund Brekke

HTTPS://WWW.NTNU.EDU/EMPLOYEES/EDMUND.BREKKE

First edition, August 13, 2019

Contents

1. Overview of tracking, navigation and SLAM

Wikipedia defines sensor fusion as *combining of sensory data or data derived from disparate sources such that the resulting information has less uncertainty than would be possible when these sources were used individually*. Depending on the application, this entails different levels of interpretation of the data. Sensor fusion is closely related to *state estimation*, as the goal is typically to estimate a state vector for an underlying dynamical system based on noisy measurements.

In the context of state estimation, there are two fundamental approaches to sensor fusion. On the one hand, it is possible to process measurements from the different sensors as they arrive, updating the state estimate every time this happens. This is known as *measurement level fusion*. On the other hand, it is possible to maintain different estimates for the different sensors, and then fuse them into a combined estimate. This latter approach is in the context of target tracking known as *track-to-track fusion*. In this book, the main focus will be on the former approach. Therefore, this is not really a book about sensor fusion, but rather about state estimation. However, the estimation problems that dominate in typical sensor fusion applications have many things in common which distinguish them from other estimation problems, and and a solid understanding of applied state estimation is essential as a fundament for more advanced work in sensor fusion.

1.1 Target tracking, navigation and SLAM

In this book the focus will be on three sensor fusion applications: Target tracking, inertial navigation and simultaneous localization and mapping (SLAM). In target tracking the goal is to estimate the motion of one or several moving objects by means of sensors that observe these objects. Such sensors can include cameras, radar, laser scanners and sonar. In inertial navigation the goal is to estimate the motion of a moving object by means of sensors attached to that object. This includes inertial sensors such as an accelerometer and a gyroscope, and sensors such as GNSS and compass which measure location and orientation relative to an external coordinate system. In SLAM the goal is again to estimate own motion as in inertial navigation, but it is also a goal to build a map of the environment. For this reason, SLAM must also use imaging sensors such as those used in target tracking.

We shall refer to the sensors used in target tracking as *exteroceptive*, while we shall refer

to the sensors used in inertial navigation as *interoceptive*. We may also refer to exteroceptive sensors as imaging sensors. While it perhaps is an abuse of terminology to describe GPS as interoceptive, this terminology makes it easy to distinguish the two categories of sensors. The processing of interoceptive and exteroceptive data is radically different. Interoceptive data (e.g., a GPS measurement) can be fed directly to a Kalman filter, which then will estimate position and other kinematic quantities[1]. For exteroceptive data, which come as images or scans, we must first extract relevant features by means of detection and segmentation procedures. This will typically give a list of blobs, which can be reduced to point features by means of clustering procedures. To be able to estimate anything from these features, it is necessary to establish correspondences between features in consecutive scans. This is known as *data association*.

Consequently, the processing pipeline in target tracking and SLAM is considerably more complex than in inertial navigation. but it is also important to understand that solutions to the different applications are dictated by rather different requirements. Inertial navigation may seem like the simplest problem since it does not involve data association, but the requirements of accuracy and robustness are generally very strict. Therefore inertial navigation uses more sophisticated models, and clever choices of estimator design are needed. On the other hand, target tracking tends to use extremely simple motion models, since it would be futile and even dangerous to attempt to estimate more complex models. While a typical navigation system uses 16 or more dimensions in the state vector, most tracking methods use only 4 or 5 dimensions in the state vector.

In inertial navigation it is often desired to make solutions with global stability properties. Such properties provide a guarantee that the filter will not diverge. For target tracking, where all reasonable motion models are marginally stable, it is impossible to guarantee stability. Indeed, if the system is allowed to run for long enough time, divergence is bound to happen. The best safeguard against this is ensure that all plausible hypotheses for data association are considered. For this reason, the computational complexity of target tracking will typically be substantially higher than for inertial navigation.

For SLAM, data association is also an issue that must be taken seriously, but it is less critical than for target tracking. The reason for this is that SLAM typically uses hundreds and thousands of landmarks in the environment. Even if the method fails to associate a large percentage of these, the remaining landmarks are still likely to carry enough information to support estimation of own motion. However, the large number of landmarks is also something that drives computational complexity up. A key research topic in SLAM has therefore been to exploit structure inherent in the SLAM problem to enable fast state estimation.

1.2 A brief history of sensor fusion

Research in target tracking has historically been driven by military applications. The radar was developed just before and during the early phases of World War 2. Most of the tracking algorithms currently in use were developed during the cold war. These were typically designed for guiding antiaircraft artillery or antisubmarine torpedoes. Variations of these methods have dissipated into virtually any defence application where the goal is to shoot someone or something.

Target tracking found its first civilian applications in air traffic control (ATC) and as a maritime navigation aid, where it is known as automatic radar plotting aid (ARPA). The main purpose of these systems is to enable human operators to prevent collisions, by keeping track of the whereabouts of other airplanes and ships, as well as unwelcome intruders such as drones. These systems have undergone progressive steps towards autonomy. Since the Überlingen aircraft collision in 2002 pilots have been instructed to trust their automatic collision avoidance system over human instructions.

[1] Wild-point filtering should of course be used to prevent obviously erroneous measurements to do any damage.

Target tracking also plays a role in all kind of surveillance applications. This can be surveillance of secured areas where one wants to know if unwelcome intruders are passing through or loitering. It can be biological or industrial applications where one wants to follow the motion of fish, bacteria, insects, etc. Tracking methods can also be used to keep track of space debris that poses a threat to satellites or astronauts.

In the same way as for target tracking, early applications of inertial navigation also emerged in World War 2 due to military needs. Long-range missiles such as the German V2 rockets needed a navigation system to hit their targets. The development of inertial navigation continued with the more advanced missiles developed during the cold war. At the same time, the space race between the United States and the Soviet Union also led to the usage of advanced navigation technology in spacecraft and satellites. The moonlanding is often credited as the first application of the Kalman filter. Since then, inertial navigation has become commonplace technology in all kinds of vehicles, and also electrical gadgets such as smartphones. The next stepping stone was the introduction of Global Navigation Satellite Systems (GNSS), including the Global Positioning System (GPS), which since 1995 has since then provided a global reference frame for navigation systems, so that location anywhere on the earth can be estimated with an accuracy of a handful of meters, or possibly centimeters if differential techniques are used.

However, access to satellites is limited in many situations, such as for underwater vehicles or indoor robotics. This limitation has since the late 90s been a major driving force for the development of SLAM. Using SLAM, a robot can estimate its own location relative to its local environments, such as the rooms in a building. SLAM also has roots in computer vision, virtual reality and 3-D reconstruction.

The primary motivation behind the course TTK4250 Sensor Fusion is current research on autonomous vehicles. Since the DARPA Grand Challenge was won by the driverless car Stanley in 2005 it has been clear that autonomous cars can successfully cope with fairly complex environments, and the automotive industry has since then been putting huge efforts into the development of autonomous cars. Both target tracking, inertial navigation and SLAM are essential components of the software for autonomous cars. Norway does not have a large automotive industry, but its maritime industry is world-leading. Increased levels of autonomy are expected to have a huge impact on this industry in the near future.

A secondary motivation behind TTK4250 Sensor Fusion is that estimation in general plays an increasing role both in cybernetics and related disciplines. A general trend is that as simpler problems are being solved, researchers and innovators move on to tackle more challenging problems. The simpler problems are in this context those that can be solved by feedback control alone, possibly including some filtering or observers to process sensor data. The more challenging problems are those that involve a more complex interpretation of the data, and where the control system consequently must perform some kind of reasoning in the presence of uncertainty. The fundamental tools such as Kalman filters, particle filters, Gaussian mixtures and graphical models appear again and again in such applications. The more general management of uncertainty is highly relevant in several other fields, including subsurface imaging for oil and gas exploration, medical imaging, meterology and climate science, system identification of physical and biological processes and financial markets, to mention a few areas.

1.3 Modeling - important or not so important?

Imagine that you hear the roar of an F16 somewhere to the west. Soon after you hear it right above you. Instinctively, you then try to spot it to your east, where you already can see it fly into the horizon. In this process, you utilize several models. First, you have a *plant model*, also known as process model or kinematic prior, which tells you that the plane probably continues more or less in the same direction as it moved earlier. Second, you have a *measurement model*, which in this

particular example would account for the fact that the airplane probably was in the direction from which the sound was emitted when the sound was emitted, and which also would account for the fact that the speed of sound is not that much higher than the speed of the airplane.

It is fairly evident that models have a central role to play in sensor fusion. For the choice of models, the Einstein maxim applies: "As simple as possible, but not simpler". Theoretically, we could make a model of the F16 plane which accounts for its motion in 6 degrees of freedom, including various dynamical forces. Unfortunately, there is no way that we can estimate all the state variables of such a model. Not only would it be meaningless to use such a model, but it would most certainly be quite disastrous, as the inevitable failures to estimate some of the states very soon will cause the track to diverge significantly from the truth. In contrast, most tracking methods use models assuming nearly constant velocity (the CV model) or nearly constant turn rate (the CT model).

The modeling task does not necessarily stop with such kinematic models, e.g., the kind of models that are used in a Kalman filter. If you see several airplanes instead of only one, then you are forced to evaluate which of these is that particular F16 plane. If you do not see the plane again, perhaps you will consider the possibility that the roar was from a meteorite fall from the sky instead. However, you are in possession of some vague prior knowledge about how frequently F16s and meteorites fly above your head, and this knowledge would probably make you extremely reluctant to accept this alternative explanation, and you would still search for the F16 in the sky. Especially in target tracking, this kind of knowledge is often encoded in models for target existence, false alarms and misdetections. Even if all these models on their own are very simple, the overall tracking problem, where all the models are combined, will quickly become rather complex.

It is important that models make sense and exhibit sufficient degree of realism, especially when expanding modeling from basic kinematic properties to more exotic properties. Sometimes one may discover that an estimation method that in theory should work extremely well, fails in any realistic scenario because its modeling assumptions are entirely unrealistic. Also, adapting a tested and tried method to a new problem may fail to be successful if that problem has characteristics which violate the assumptions of that method.

Machine learning methods are often viewed as an alternative to model-based methods. This is a valid point of view, although with two important caveats. It cannot be emphasized strongly enough that the term machine learning is a wide umbrella that covers several loosely related methods. As the first caveat, so-called probabilistic machine learning methods do involve explicit models, which are to be learned from the data. As the second caveat, artificial neural networks are fundamentally nonlinear regression engines which encode relationships between inputs and outputs in a black box fashion. In principle there are no limits to the relationships that such an engine can learn. It could for example learn the relationship between GPS and IMU measurements and the true state vector in a navigation system. But this relationship can be viewed as an implicit representation of the kinematic model. The question then is whether the learned model will give better predictive power than a model specified manually from prior knowledge. For complex problems such as image recognition this is probably the case, while it is much more questionable for, e.g., inertial navigation using GPS and IMU.

1.4 The deterministic and probabilistic paradigms

Uncertainty can be dealt with in many ways. First of all, we may ignore it alltogether. Many classical estimation techniques exist which do not actually utilize models of uncertainty. In standard unweighted least squares estimation one simply aims to find the underlying parameter that minimizes the average error that results between observations and predicted observations. As another example, to match two images one may extract some features in both images, and attempt to find a transformation that makes as many of the features as possible coincide.

State estimation is often referred to as filtering, and state estimators are often known as filters (e.g., the Kalman filter). Filtering is a fairly general concept. One can design filters which do not correspond to a meaningful estimators, and one can design filters that estimate a state without any explicit treatment of uncertainty. Instead, a filter is typically designed by specifying requirements on the spectral properties of the estimation error. For example, in dynamical positioning the goal is to estimate long term motion of a vessel, while short term disturbances due to sea waves should be ignored. This way of thinking leads naturally to the theory of linear and nonlinear observers, whose tuning parameters (gains) are tuned according to the desired response time of the observer. The designer of the observer will typically present a stability proof to convince that the observer can be relied upon.

The conventional approach in sensor fusion is, however, to quantify uncertainty by means of probabilities. This is quite natural, as probability has been the established language of uncertainty since the time of Laplace. There are several reasons why we would like to quantify uncertainty, as opposed to merely coping with uncertainty. First of all, established estimation algorithms such as the Kalman filter have been developed in a probabilistic context. Furthermore, quantifying the uncertainty is useful to help us with interpreting the results. Returning to the F16 example, having an uncertainty estimate makes it possible for you to decide how much you should move your gaze around the spot where you expect to see it. As a third reason, quantified uncertainty enables us to make precise rules for how to weigh different pieces of evidence against each other in a (multi-) sensor fusion system.

Such rules may be perceived as a straightjacket or as the beams of a house construction. This book will attempt to convince you towards the latter perception. It is extremely important to have sensible structures to rely upon when designing complex sensor fusion systems. At the core of these structures we often find Bayes' rule, which is a very powerful tool, because it governs how information changes in the presence of new evidence. Proving that a sensor fusion method obeys Bayes' rule is in some sense the equivalent of stability proofs in control theory. The virtue of a stability proof is not necessarily that the system will converge to zero error. If noise continuously enter the system, that will not happen. The stability proof may, however, ensure that the system will not do anything crazy. Similarly, if a sensor fusion method is faithful to Bayes, we can trust it to behave in a rational manner.

1.5 References and rationale

Several excellent books on estimation and sensor fusion exist. Why was it then necessary to write a textbook from scratch for the course TTK4250? The answer has to do with the limitations of how much knowledge a student can be expected to digest during a 7.5 study points course. In TTK4250 the goal is to reach sufficient comprehension of state-of-the-art methods for tracking, navigation and SLAM to be able to use these from day one in research problems in the 5th year specialization projects and MSc theses. Furthermore, this goal is to be achieved from a foundation consisting of little more than a basic probability course and a basic Kalman filtering and estimation course (i.e., linear systems theory). There exists no textbook today that covers this entire span. Nevertheless, I am convinced that it is possible to cover this in a one-semester course, if there is a strong focus on the core methods and their theoretical foundations, and less focus on general theory or the many possible variations of the methods.

If this course was to use an existing textbook, the most natural candidates would have been the 2001 book by Bar-Shalom, Li & Kirubarajan [2] or the 2012 book by Gustafsson [35]. Both cover the fundamentals of probability and estimation extremely well, and include authoritative in-depth treatments of Kalman filtering, several nonlinear methods (such as EKF) and multiple model estimation. Any student who wants to do serious work in sensor fusion should own a copy of at least one of these. There are, however, two reasons why I have chosen not to use any of these as

curriculum for TTK4250. Both books are very rich on details, and I am concerned that the amount of details will simply be too much for many students to digest. Furthermore, neither book covers data association, which in my opinion is one of the most important topics in sensor fusion.

Fundamental probability and estimation theory is covered in Chapters 2 and 3, but only to a limited extent. For more details the reader is referred to Kay's book [42] from 1993 and to Papoulis & Pillai's textbook on probability and stochastic processes from 2002 [60]. Again, any serious sensor fusion student should be familiar with these books. While [42] is very readable and takes a discrete-time approach, [60] is the definitive reference on the theory of continuous-time stochastic processes.

For target tracking, I am first and foremost inspired by Bar-Shalom and Li, represented by the book [1], which was written in 1995. A more expanded version of this book came in 2011 with the title title "Tracking and Data Fusion: A Handbook of Algorithms" [4]. The most comprehensive textbook on target tracking that exists is probably [7]. It is indispensable for diving into the building blocks of MHT implementations, but cannot be used as an authoritative source on the state-of-the-art with its 20 years of age. To become familiar with the modern paradigm in target tracking the reader would have to supplement the mentioned books with Challa [17] and Mahler [50] (or the research articles that these books build upon). In [17] methods that use the concept of existence probability are described. The book is to an even larger extent than [4] written as a handbook of algorithms. In [50], the focus is on Mahler's theory of random finite sets. While Mahler has made significant efforts to present this material in a pedagogical manner, the material itself is far too demanding to be used in a basic course on sensor fusion.

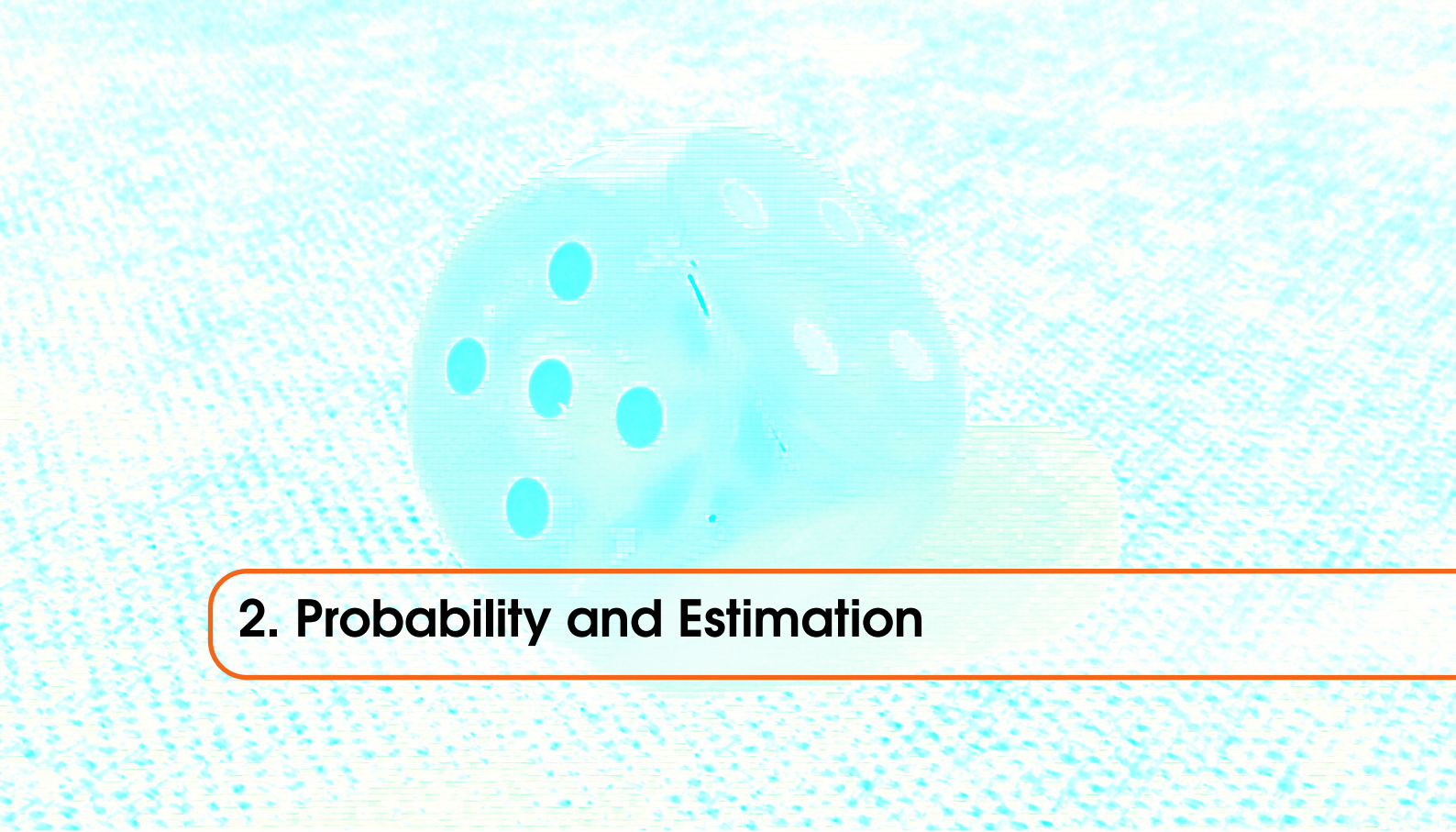
As for navigation, established textbooks include the books by Groves [33] and Titterton & Weston [73]. These delve much deeper into the fundamentals of GNSS technology etc. than we would like to do in this course, while their exposition of the error state formulation of the Kalman filter is much less systematic. Inertial navigation has previously been taught in the course TTK5 using Vik's book [77]. Again, the main shortcoming is no systematic exposition of the error state formulation. This is, however, available in Sola's text [70], which forms the basis for our treatment of inertial navigation.

There exists no textbook on SLAM that is up to date with the significant developments that have found place during the last 10 years. The closest one gets to a canonical textbook on SLAM remains Thrun's book [72] which is from 2006. It gives a very detailed treatment of EKF-SLAM and particle filtering SLAM. Recursive SLAM methods, together with the related topic of localization, are also covered in [35]. While not primarily about SLAM, Barfoot's book [5] may be very relevant for anyone working on SLAM and navigation.

1.6 Acknowledgement

This book is largely inspired by my experience in supervising MSc and PhD students working on autonomous ship projects during the last 5 years. All of these students can be found on the website `folk.ntnu.no/edmundfo/autoseastudents/autoseastudents.html`. In particular, I would like to express special appreciation to Lars-Christian Tokle, who has contributed ideas, proofreading and some of the central proofs, and Michael Ernesto Lopez who has made several of the TiKZ-coded figures. The book has been written using Texmaker and the Legrand Orange Book template (`www.latextemplates.com/template/the-legrand-orange-book`). During the writing I upgraded my macbook to a 2018 model with butterfly keys. I blame all typos on the butterfly keys.

Finally the reader should be aware that this is still work in progress. As of 14th of August, only the first three chapters are completed and will be released on Blackboard. Additional chapters will be released as soon as they are completed.



2. Probability and Estimation

2.1 Foundations of probability theory

Probability theory is about assigning numbers to *events*. These numbers, known as probabilities, tell us something about how likely these events are to happen. This intuitive notion of probability can be expressed more precisely in terms of three axioms. Let the letter E refer to any event, and let Ω be the outcome space, that is, the collection of all possible events. We use the notation $\Pr\{\cdot\}$ to signify probability of the events. The axioms of probability can then be formulated as follows

1. $\Pr\{E\} \in \mathbb{R}$, $\Pr\{E\} \geq 0$.
2. $\Pr\{\Omega\} = 1$.
3. For any sequence of disjoint events $E_1, \ldots, E_n$ we have $\Pr\{\bigcup_{i=1}^{n} E_i\} = \sum_{i=1}^{n} \Pr\{E_i\}$.

The first axiom then tells us that the probability of any event is a non-negative real number. The second axiom tells us that the probability of the entire outcome space is one. The third axiom tells us that the probabilities of disjoint events can be added in order to find the probability that any of the events considered should happen.

In addition to the axioms, probability theory also relies on some fundamental definitions, which provide a language for how different events can be related to each other.

Definition 2.1.1 — Conditional probability. Given two events A and B, the conditional probability of A given B is

$$\Pr\{A|B\} = \frac{\Pr\{A \cap B\}}{\Pr\{B\}} \tag{2.1}$$

Definition 2.1.2 — Independence. We say that two events A and B are independent if

$$\Pr\{A \cap B\} = \Pr\{A\}\Pr\{B\} \tag{2.2}$$

Equipped with these definitions, we can establish some basic results in probability theory. The following two results will play a key role in very many of the derivations in the subsequent chapters.

Theorem 2.1.1 — The total probability theorem. Let $\{B_n\} = \{B_1, B_2, B_3, \ldots\}$ be a countable partition of the outcome space, and let A be some event. Then the following is true:

$$\Pr\{A\} = \sum_n \Pr\{A|B_n\}\Pr\{B_n\}.$$

Proof. The probability of A is the same as the probability of the union $\bigcup_n (A \cap B_n)$ insofar as $\{B_n\}$ is a partition of the outcome space. Thus

$$\Pr\{A\} = \sum_n \Pr\{A \cap B_n\} = \sum_n \Pr\{A|B_n\}\Pr\{B_n\}$$

where the first equality follows from axiom number 3 and the last equality follows from the definition of conditional probability. ■

Theorem 2.1.2 — Bayes' rule. For any two events A and B, the following is true:

$$\Pr\{A|B\} = \frac{\Pr\{B|A\}\Pr\{A\}}{\Pr\{B\}}.$$

Proof. According to the definition of conditional probability the joint probability of A and B is $\Pr\{A \cap B\} = \Pr\{A|B\}\Pr\{B\} = \Pr\{B|A\}\Pr\{A\}$. Bayes' rule follows from dividing by $\Pr\{B\}$ on both sides. ■

2.2 Random variables and probability distributions

It would be very cumbersome if we always were to work explicitly with events. In probability theory and its applications we are typically interested in the behavior of quantities that are of a random nature. By the phrase "of a random nature" we mean that the quantity is uncertain, and that our knowledge about it is modeled by means of probability theory. Such a quantity could for example be the speed of an airplane. In a probabilistic model we would have some probability that it is smaller than $100\,\mathrm{m/s}$, while the probability that it is larger than or equal to $100\,\mathrm{m/s}$ would be one minus that probability. Statements such as "the velocity is less than $100\,\mathrm{m/s}$" correspond to the events in Section 2.1.

2.2.1 Random variables and probability measures

Random variables are in the mathematical approach to probability theory defined as functions from an abstract outcome space to a more concrete outcome space such as $\mathbb{R}^n$. For any element in the abstract space the random variable attains a certain element in the concrete space, known as the *realization* of the random variable. The formal definition of a random variable is based on measure theory, which is one of the core foundations of mathematical analysis (other foundations being algebra and topology). The conceptual framework revolves around the concept of a *probability space*, which again builds on the concepts of σ*-algebras* and *measures*.

Definition 2.2.1 — σ-algebra. A σ-algebra $\mathscr{F}$ on a set Ω is a collection of subsets of Ω such that

- Ω itself is a member of $\mathscr{F}$.
- If the subset $A \subseteq \Omega$ is a member of $\mathscr{F}$, then its complement $\Omega \backslash A$ is also a member of $\mathscr{F}$.
- If A_1, A_2, A_3, etc are members of $\mathscr{F}$, then the union $\bigcup_n A_n$ is also a member of $\mathscr{F}$.

In probability theory, the elements of the σ-algebra plays the role of meaningful events. We can illustrate this with a simple example.

■ **Example 2.1 — Throw of a dice.** The outcome space when we throw a dice is $\Omega = \{1,2,3,4,5,6\}$. All the possible subsets of these 6 elements constitute a corresponding σ-algebra:

$$\mathscr{F} = \{\{1\},\{2\},\ldots,\{1,2\},\ldots,\{1,2,3,4,5,6\}\}.$$

We see that for any outcome at least one, and possibly several members of the σ-algebra will be active. ■

Measures are functions from σ-algebras to the real numbers. In other words, a measure is a systematic way of assigning numbers to subsets. In probability theory we are only concerned with a special class of measures known as probability measures.

Definition 2.2.2 — Probability measure. A probability measure P on the σ-algebra $\mathscr{F}$ is a function from $\mathscr{F}$ to the unit interval $[0,1]$ that obeys the three axioms of probability.

We see that the probability measure by definition obeys the fundamental axioms of probability that we introduced in the beginning of Section 2.1. We refer to the combination of the outcome space, the σ-algebra, and the probability measure as a probability space.

Definition 2.2.3 — Probability space. A probability space is a triplet $(\Omega,\mathscr{F},P)$ where

- Ω is a space of possible events.
- $\mathscr{F}$ is a σ-algebra on Ω.
- $P : \mathscr{F} \to [0,1]$ is a probability measure.

Definition 2.2.4 — Random variable. A random variable X is a function from Ω into another space $\mathbb{O}$, which we henceforth shall know as the outcome space.

The outcome space $\mathbb{O}$ can be many different things. In game of dice it will consist of the possible faces of a dice, which can be represented as $\{1,2,\ldots,6\}$. Most typically, $\mathbb{O}$ will be the space of real numbers $\mathbb{R}$, or it may the n-dimensional space of real-valued vectors $\mathbb{R}^n$. In the latter case, we often talk about random vectors instead of random variables. The random variable X connects the abstract space with the outcome space $\mathbb{O}$, so that probability becomes distributed on $\mathbb{O}$. The probability that the value of X is a member of B, where $B \subseteq \mathbb{O}$, is given by $P(X^{-1}(B))$, i.e., by summing the probabilities of all elements $\omega \in \Omega$ for which $X(\omega)$ ends up in B.

Why is it useful to introduce the abstract space Ω in addition to the more tangible outcome space $\mathbb{O}$? At least two answers can be provided. First, the definition of random variables as functions provides a clear mechanism for distinguishing random variables from their realizations. While the difference is intuitively easy to grasp, it is also nice to have a precise mathematical distinction. Second, the tangible outcome space $\mathbb{O}$ is not necessarily as neat and tidy as $\mathbb{R}^n$. For example, we shall in Chapter 10 let $\mathbb{O}$ consist of all finite subsets of $\mathbb{R}^n$. As another example, $\mathbb{O}$ could be a closed manifold such as $SO(3)$ or the surface of the earth. In such cases it may not be obvious how probability mass should be distributed in $\mathbb{O}$. By defining the random variables as functions we get a mechanism for mapping the probability mass from a space where probabilities by definition make sense to this more exotic space.

Definition 2.2.5 — Realization. The output of the random variable X for a particular $\omega \in \Omega$ is called a realization of X. We can write this as $x = X(\omega)$.

The concepts of a random variable and its realization are often confused, accidentally or deliberately. Often the same notation is used for both. This is often the only sensible thing to do, as notation would quickly become very complicated if we were to distinguish random variables and their realizations all the time. Nevertheless, the distinction should be remembered, as it sometimes can be quite important.

The abstract measure $P(\cdot)$ induces a probability measure on the tangible outcome space $\mathbb{O}$. We can relate this measure to probabilities as follows:

Definition 2.2.6 — Probability measure of a random variable. The probability measure of X is a function $\beta_X(\cdot)$ from subsets of $\mathbb{O}$ to $[0,1]$ so that

$$\beta_X(S) = \Pr\{X \in S\} = P(X^{-1}(S)). \tag{2.3}$$

2.2.2 Probability distributions

The main issue with probability measures, and the equivalent formulation in terms of primitive events in Section 2.1, is that these representations are highly redundant. To see this, and make the conceptual leap over to more useful representations, let us look at scalar continuous-valued random variables, i.e., let $\mathbb{O} = \mathbb{R}$. We do not need to assign a probability value to all possible subsets of $\mathbb{R}$ to specify the random properties of the random variable X. It would suffice to assign a probability value to all subsets of the form $(-\infty, x)$ where x is a real number, i.e., a possible realization of X. Or it would suffice to define a function whose integrals from $-\infty$ to x yields these values. These two viewpoints lead to the concept of cumulative distribution functions and probability density functions.

Definition 2.2.7 — Cumulative distribution function (cdf). The cdf, denoted $P(x)$, of $X \in \mathbb{R}$ is the probability $\Pr\{X < x\}$.

Definition 2.2.8 — Probability density function (pdf). The pdf of the scalar random variable X is the derivative

$$p(x) = \frac{\partial P(x)}{\partial x}.$$

Generally, the term probability distribution is used to mean the same as a pdf. It must be emphasized that *pdf's and probabilities are two different things.* We get probabilities when we integrate a pdf over a subset of the outcome space.

For discrete-valued random variables (e.g., throwing a dice) we do not normally talk about the pdf, but rather about the point mass function (pmf), which is given by $p(x) = \Pr\{X = x\}$. We can obtain the cdf in the same manner as for continuous-valued random variables by replacing integration with summation. Notice that such expressions such as $\Pr\{X = x\}$ in general are meaningless for continuous-valued random variables: The probability of X attaining a particular value will be zero if X is distributed over a continuous space.

The extension of cdfs and pdfs to vector-valued random variables is fairly straightforward. For example, if $Z = [X,Y]^\mathsf{T}$ for scalar X and Y, then the cdf is defined as

$$P_Z(z) = P_{X,Y}(x,y) = \Pr\{X \leq x, Y \leq y\}$$

and the pdf is found as

$$p_Z(z) = p_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} P_{X,Y}(x,y) = \frac{\partial^2}{\partial y \partial x} P_{X,Y}(x,y).$$

It must again be emphasized that the abstract machinery of Section 2.2.1 is much more general than random variables in $\mathbb{R}^n$. More care is needed if one aims to extend the constructions of cdfs and pdfs to exotic state spaces such as closed manifolds.

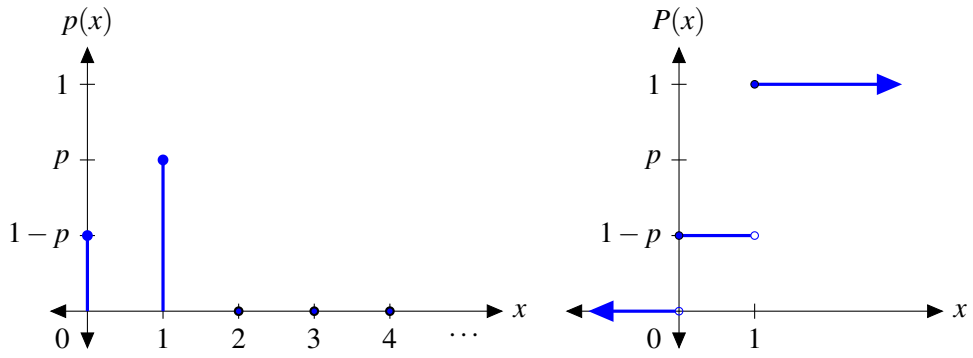
2.2.3 Examples of probability distributions

It is time to look at some examples of random variables that play important roles in sensor fusion. We shall first look at a handful of discrete examples.

■ **Example 2.2 — Bernoulli random variable.** A Bernoulli random variable with parameter r has a binary outcome space: $E = \{0,1\}$, and its probability distribution is given by

$$p(x) = \begin{cases} 1-r & \text{if } x = 0 \\ r & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

We write $p(x) = \text{Bernoulli}(x;r)$ to signify this distribution. ■



(a) PMF. (b) CDF.

Figure 2.1: Example of Bernoulli distribution.

■ **Example 2.3 — Binomial random variable.** A binomial random variable with parameters $r \in [0,1]$ and $n \in \mathbb{N}$ has outcome space $\{0,\ldots,n\}$ and its probability distribution is given by

$$p(x) = \binom{n}{x} r^x (1-r)^{n-x}. \tag{2.5}$$

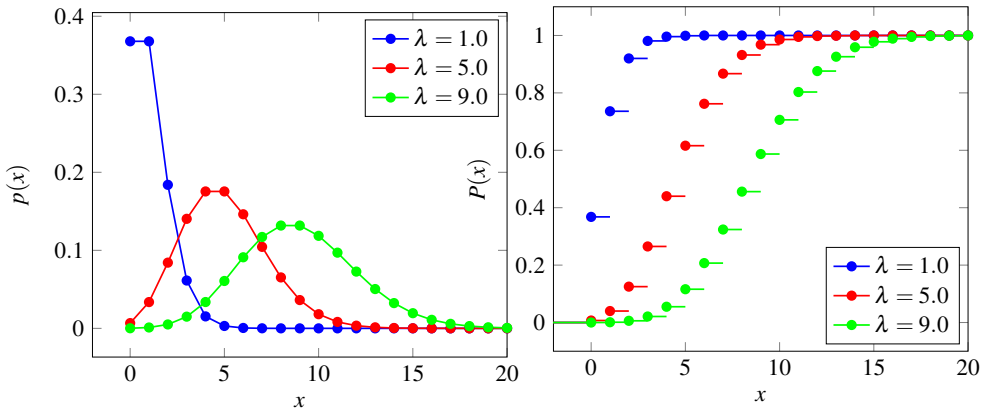
■

■ **Example 2.4 — Poisson random variable.** A Poisson random variable with parameter λ has the countable (that means infinite, but discrete) outcome space $\{0,1,2,3,\ldots\}$ and its probability distribution is given by

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{2.6}$$

We write $p(x) = \text{Poisson}(x;\lambda)$ to signify this distribution. ■

Bernoulli random variables are often used to model whether a target exists or does not exist, or whether a target is detected or not detected. Poisson random variables are often used to model how many false alarms there are in a radar scan. The binomial random variable can be viewed as a generalization of the Bernoulli random variable when there is more than one yes-no question, e.g., two potential targets, or two sensor cells which may or may not contain an observation. According to the Poisson limit theorem, also known as the law of rare events, the binomial distribution can be approximated by the Poisson distribution if n tends towards infinity and r tends towards zero in such a manner that the product nr tends towards λ.



(a) PMF (b) CDF

Figure 2.2: Poisson distribution.

As for continuous random variables, the simplest possible example is the uniform distribution, which for instance often is used in target tracking to model the spatial distributions of false alarms or newborn targets.

■ **Example 2.5 — Uniform random variable.** A uniformly distributed random variable X on the interval $[a,b]$ has the pdf

$$p(x) = \text{Uniform}(x;[a,b]) = \frac{1}{b-a}\chi_{[a,b]}(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise.} \end{cases} \tag{2.7}$$

Its cdf can be written as

$$p(x) = \begin{cases} 0 & \text{if if } x \le a \\ \frac{x-a}{b-a} & \text{if } a \le x \le b \\ 1 & \text{if } x \ge a. \end{cases} \tag{2.8}$$

■

By far the most important class of continuous random variable are Gaussian random variables. In a similar manner as for uniform random variables, the pdf of a Gaussian random variable is given by two parameters. It is possible to parameterize the pdf in different ways.

■ **Example 2.6 — Gaussian random variable.** A Gaussian random variable with expectation μ and variance σ^2 has the pdf

$$p(x) = \mathcal{N}(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{2.9}$$

This is known as the moment-based parameterization. Another parameterization is the canonical parameterization. If we define $\eta = \mu/\sigma^2$ and $\lambda = 1/\sigma^2$, then the same pdf can also be expressed as

$$p(x) = \exp\left(\alpha + \eta x - \frac{1}{2}\lambda x^2\right) \quad \text{where} \quad \alpha = -\frac{1}{2}\left(\ln(2\pi) - \ln(\lambda) + \eta^2/\lambda\right). \tag{2.10}$$

The cdf of the Gaussian does not exist in closed form. It is typically expressed in terms of the so-called error function according to

$$P(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right] \quad \text{where} \quad \text{erf}(x) = \frac{1}{\sqrt{\pi}}\int_0^x \exp(-t^2)\mathrm{d}t. \tag{2.11}$$

■

The Gaussian distribution is the most well-known example of a distribution from the so-called exponential family. Another important member of this family is the Gamma distribution.

■ **Example 2.7 — Gamma and χ^2 random variables.** A Gamma random variable with shape parameter k and scale parameter θ has the pdf

$$p(x) = \text{Gamma}(x; k, \theta) = \frac{x^{k-1}\exp(-x/\theta)}{\theta^k \Gamma(k)} \tag{2.12}$$

Again different parameterizations are possible, but the parameterization in terms of shape and scale is sufficient for us. Several other exponential-family distributions are intimately linked to the Gamma distribution. Both the Rayleigh distribution, the exponential distribution and the χ^2-distribution are special cases of the Gamma distribution. The exponential distribution $p(x) = \lambda e^{\lambda t}$ arises as a special case of (2.12) when $k = 1$ and $\theta = 1/\lambda$. The Rayleigh distribution $p(x) = \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)}$ arises as a special case when $k = 2$ and $\theta = 2\sigma^2$. As for the χ^2 distribution, we say that X is χ^2 distributed with n degrees of freedom if

$$p(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} \exp\left(-\frac{x}{2}\right). \tag{2.13}$$

We can see that (2.13) is a special case of (2.12) when the scale parameter is equal to 2 and the shape parameter is equal to $\frac{n}{2}$. For general Gamma and χ^2 random variables the cdf is again non-analytic, while closed-form expressions are easily found for Rayleigh and exponential RVs. Nevertheless, it is often important to find such cdf values, and to solve for x when a cdf value is given, and this can easily be done with Matlab functions such as `chi2cdf` and `chi2inv`, respectively. ■

2.2.4 Sampling from probability distributions

To play with random variables, it is very useful to be able to simulate them. Programming languages such as Matlab comes with support or add-on packages which allow the user to draw samples from well-known probability distributions without putting much thought into how the simulation is done. For examples, to draw a 4×1 vector of independent samples from $\mathcal{N}(x; 0, 1)$ we can use the Matlab command

```
randn(4,1).
```

More complicated random variables can often be simulated by exploiting their relationships to other random variables. For example, we shall later see (Example 2.11 on page 27) that the absolute value of such a Gaussian random variable is a χ^2 distributed random variable with one degree of freedom. Thus, we have a means of simulating such χ^2 distributed random variables.

If we are not so lucky to have such a relationship to exploit, we can still use *cdf inversion*, also known as inverse transform sampling or the Smirnov transform. Consider the problem of generating N independent samples of a random variable X with pdf $p_X(x)$. In this approach, we draw a random N numbers u^i from the uniform distribution $\text{Uniform}(u; [0, 1])$. The we look at the cdf $P_X(x)$. Every sample x^i is then found as the value of x^i such that

$$P_X(x^i) = u^i. \tag{2.14}$$

In other words, we find the samples as $x^i = P_X^{-1}(u^i)$. We see that this approach relies fundamentally on our ability to invert the cdf. This may or may not be possible to do in closed form. However, even if no closed-form solution exist, approximations may be good enough. For example, we can approximate the cdf by "segmented" functions, in the shape of a staircase or a C^0 collection of straight lines.

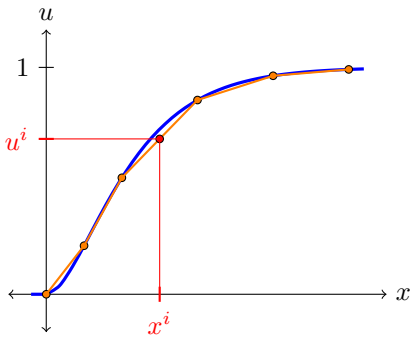


Figure 2.3: Illustration of how the Smirnov transform can be used to sample a random variable even if its inverse cdf has no analytical expression.

2.3 Moments

In many cases, we are not primarily interested in the full pdf of a random variable, but only in certain descriptive measures which give us more tangible information about how the random variable behaves. Among such measures are moments, which are expectations of power of the random variable. The two most well known moments are expectation and variance.

The expectation and variance of a scalar-valued random variable X with realization x are given by the integrals

$$E[X] = \int xp(x)\mathrm{d}x \tag{2.15}$$

$$\mathrm{Var}[X] = E[(X - E[X])^2] = \int (x - E[X])^2 p(x)\mathrm{d}x. \tag{2.16}$$

The notations $\mu = E[X]$ and $\sigma = \sqrt{\mathrm{Var}[X]}$ are commonly used. The generalizations for a vector-valued random variable X with realization $\mathbf{x}$ are given by

$$E[X] = \int \mathbf{x}p(\mathbf{x})\mathrm{d}\mathbf{x} \tag{2.17}$$

$$\mathrm{Var}[X] = E[(X - E[X])(X - E[X])^\mathsf{T}] = \int (\mathbf{x} - E[X])(\mathbf{x} - E[X])^\mathsf{T} p(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{2.18}$$

In the vector case $\mathrm{Var}[X]$ becomes a matrix, and not merely a number. It is known as the covariance matrix and consists of elements of the form

$$\mathrm{Cov}[X_1, X_2] = E[(X_1 - E[X_1])(X_2 - E[X_2])^\mathsf{T}] = \int (x_1 - E[X_1])(x_2 - E[X_2])p(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2. \tag{2.19}$$

For discrete random variables the integrals are replaced by sums. The expectation and the variance are closely related to the sample mean and the sample variance, which can be calculated for a collection of N samples according to

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \tag{2.20}$$

$$\mathbf{P} = \frac{1}{N-1}\sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\mathsf{T} \tag{2.21}$$

If the samples are drawn from the distribution $p(x)$, then the mean will converge towards its expectation as the number of samples goes towards infinity. Notice that these sample-dependent quantities in contrast to the expectation and variance are random variables. The distribution of the sample mean can in some cases be calculated analytically, while in other cases it may be intractable. However, as the number of samples goes towards infinity we are guaranteed some regularity thanks to the Central Limit Theorem (CLT).

Theorem 2.3.1 — Central Limit Theorem, Lindeberg-Lévy version. Let $X_1, X_2, \ldots X_n$ be a sequence of i.i.d. random variables with expectation $E[X_i] = \mu$ and $\text{Var}[X_i] = \sigma^2 < \infty$. Let

$$S_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then, as n approaches infinity, the distribution of the random variable $\sqrt{n}(S_n - \mu)$ converges to $\mathcal{N}(0, \sigma^2)$.

Proof. This is left as an exercise to the reader (Exercise 2.4) since it depends on generating functions, which will be introduced in Section 2.4. ■

The CLT is often used as a justification for assuming that something is Gaussian distributed. In order to assess such justifications it is important to be aware of its requirements. The random variables must be i.i.d. If one wants to analyze the sum of random variables with slightly different distributions, the CLT may break down. Also notice that $\text{Var}[X_i]$ must exist. There exist distributions for which the variance is infinity, and for such distributions the CLT cannot be used.

Since expectation is given by an integral it inherits the linearity property of the integral. Therefore, we can in general write

$$E[\mathbf{A}X + \mathbf{B}Y] = \mathbf{A}E[X] + \mathbf{B}E[Y] \tag{2.22}$$

where X and Y are random vectors and $\mathbf{A}$ and $\mathbf{B}$ are matrices. Variance involves squaring and is therefore not a linear operation. Nevertheless, the squaring operation is so benign that a similar result also can be stated for variance:

$$\text{Var}[\mathbf{A}X + \mathbf{B}Y] = \mathbf{A}\text{Var}(X)\mathbf{A}^\mathsf{T} + \mathbf{B}\text{Var}(Y)\mathbf{B}^\mathsf{T} + \mathbf{A}\text{Cov}(X,Y)\mathbf{B}^\mathsf{T} + \mathbf{B}\text{Cov}(Y,X)\mathbf{A}^\mathsf{T}. \tag{2.23}$$

We notice that variance also is a kind of expectation, which involves taking the expectation of an expectation. It can also be useful to calculate the expectation of an expectation in other contexts. The law of total expectation states that

$$E_X[X] = E_Y[E_X[X|Y]]. \tag{2.24}$$

Here we have introduced subscripts on the expectation operation to specify which random variable it applies to. If the random variables X and Y are independent, then $E_X[X|Y]$ does not depend on Y, and it follows that

$$E_{X,Y}[XY] = E_X[E_Y[XY]] = E_Y[E_X[X|Y]Y] = E_X[X]E_Y[Y]. \tag{2.25}$$

Expectation tends to have a smoothing behavior. For this reason there exist several inequalities that expectations must obey. The most famous such inequality is Jensen's inequality, which states that for an convex function $f(\cdot)$ the following must hold:

$$f(E[X]) \leq E[f(X)] \tag{2.26}$$

Such a convex function could for example be $f(x) = x^2$. Jensen's inequality can be proved by noting that the convexity of $f(\cdot)$ implies that $f(X) \geq f(E[X]) + f'(E[X])(X - E[X])$, which again implies that $f(X) \geq f(E[X])$ because $E[X - E[X])] = 0$.

We can also study moments of higher order than expectation (first-order) and variance (second-order). In particular, the fourth order moment is often used to measure how heavytailed a distribution is. By normalizing this according to the variance squared, we get the kurtosis

$$\text{Kurt}[X] = \frac{E[(X-\mu)^4]}{\sigma^4} \tag{2.27}$$

which is the standard measure for heavytailedness. For the univariate Gaussian distribution it can be shown that the kurtosis is 3. Distributions with higher kurtosis are known as leptokurtic or heavytailed, while distributions with lower kurtosis are known as platykurtic. We can also analyse how much a distribution differs from the shape of the Gaussian by evaluating its skewness, which is defined as

$$\gamma = \frac{E[(X-\mu)^3]}{\sigma^3}. \tag{2.28}$$

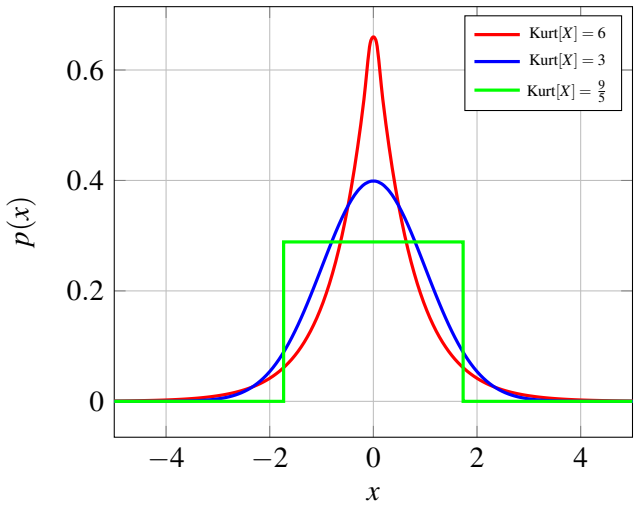


Figure 2.4: Kurtosis values for three symmetric distributions that have mean of zero and variance of one: Laplace distribution (red), normal distribution (blue) and uniform distribution (green).

A variety of information measures which describe random variables can be expressed and understood as expectations. These include the entropy

$$H[X] = E[-\ln(p(X))] \tag{2.29}$$

and the Kullback-Leibler divergence

$$D_{\text{KL}}(p,q) = E\left[-\ln\left(\frac{q(X)}{p(X)}\right)\right] \tag{2.30}$$

which provides a measure for how similar an approximating distribution $q(X)$ is to the true distribution $p(X)$. Such measures play a central role in many modern estimation methods.

2.4 Generating functions and transformations of RVs

We recall from basic control theory that sometimes it is easier to manipulate functions through their Laplace and Fourier transforms, than to work with the functions themselves. This is for example often the case when solving differential equations. In the same way, several transform-domain representations of probability distributions exist, and do often provide convenient representations. These are known as generating functions.

A generating function of a random variable is the expectation of a certain transformation of that random variable. The three most commonly encountered generating functions are known as the characteristic function, the moment-generating function and the probability-generating function. The first two are used for continuous random variables, and are in the scalar case given by

$$\Phi_X(\omega) = E_X[e^{\mathrm{i}\omega x}] = \int_{-\infty}^{\infty} p(x)e^{\mathrm{i}\omega x}\mathrm{d}x \tag{2.31}$$

$$M_X(s) = E_X[e^{sx}] = \int_{-\infty}^{\infty} p(x)e^{sx}\mathrm{d}x. \tag{2.32}$$

The latter function is used for discrete random variables, and is (again in the scalar case) given by

$$G(t) = E_X[t^X] = \sum_{n=-\infty}^{\infty} p(x_n)t^{x_n}. \tag{2.33}$$

All of these representations share three important properties:

1. The generating function determines the distribution and vice versa.
2. The generating function of a sum of independent random variables is the product of the generating functions.
3. The moments can be found by differentiating the generating function.

Property 1 should not come as a surprise. We see that $\Phi_X(\omega)$ and $M_X(s)$ work in a manner very similar to Fourier and Laplace transforms, respectively, and we know that the same principle applies to Fourier and Laplace transforms. We recognize the probability generating function as the z-transform of the pmf.

Property 2 is the most important reason why one should be familiar with generating functions: Whenever a random variable is a sum of two or more other random variables with known distributions, generating functions is a convenient tool to find the distribution of the random variable in question. Property 3 tends to be the rationale under which generating functions are presented in basic statistics courses. It is also important: It can be much more convenient to find moments by derivation than through the alternative route of integration.

■ **Example 2.8 — Sum of Bernoulli and Poisson.** Let $X \sim \text{Bernoulli}(x;p)$ and let $Y \sim \text{Poisson}(y;\lambda)$, and let $S = X + Y$. What is the probability distribution of S?
Solution: The probability generating functions of X and Y can be found as

$$G_X(t) = 1 - r + rt \tag{2.34}$$

$$G_Y(t) = \exp(\lambda(t-1)). \tag{2.35}$$

The probability generating function of S is therefore

$$G_S(t) = (1-r+rt)\exp(\lambda(t-1)) = (1-r)\exp(\lambda(t-1)) + rt\exp(\lambda(t-1)) \tag{2.36}$$

We get the probabilities of S by differentiating the probability generating function. Let us first notice that

$$\frac{\mathrm{d}^k}{\mathrm{d}t^k}\left[te^{\lambda t}\right] = \frac{\mathrm{d}^{k-1}}{\mathrm{d}t^{k-1}}\left[e^{\lambda t} + \lambda te^{\lambda t}\right] = \ldots = k\lambda^{k-1}e^{\lambda t} + \lambda^k te^{\lambda t}. \tag{2.37}$$

Based on this we find the probability that $S = k$ according to

$$\begin{aligned} p(k) &= \frac{1}{k!}\frac{\mathrm{d}^k}{\mathrm{d}t^k}G_S(t)\bigg|_{t=0} \\ &= \frac{1}{k!}\left((1-r)e^{-\lambda}\frac{\mathrm{d}^k}{\mathrm{d}t^k}\left[e^{\lambda t}\right] + re^{-\lambda}\frac{\mathrm{d}^k}{\mathrm{d}t^k}\left[te^{\lambda t}\right]\right)\bigg|_{t=0} \\ &= \frac{1}{k!}\left(\lambda^k(1-r)e^{-\lambda} + rk\lambda^{k-1}e^{-\lambda}\right). \end{aligned} \tag{2.38}$$

■

■ **Example 2.9 — Sum of Exponentials.** Let X_i, $i = 1,\ldots,N$ be N i.i.d. exponential RVs with parameter λ, and define $Y = \sum_{i=1}^{N} X_i$. What is the probability distribution of Y?
Solution: The moment-generating function for each of the i.i.d. exponentials is

$$M_X(s) = \lambda\int_0^\infty e^{(s-\lambda)x}\mathrm{d}x = \frac{\lambda}{\lambda - s} \quad \text{if} \quad s < \lambda. \tag{2.39}$$

According to Property 2 of the generating function this implies that

$$M_Y(s) = \left(\frac{\lambda}{\lambda - s}\right)^N. \tag{2.40}$$

If we were to calculate the inverse Laplace transformation of such an expression, we would end up with terms looking something like $x^N\exp(-\lambda x)$. Since this is precisely what we have in the Gamma function, let us check its moment-generating function:

$$M_Y(s) = \frac{1}{\theta^k\Gamma(k)}\int_0^\infty x^{k-1}\exp\left(-x\left(\frac{1}{\theta} - s\right)\right)\mathrm{d}x = \frac{1}{\theta^k\Gamma(k)}\Gamma(k)\left(\frac{1}{\frac{1}{\theta} - s}\right)^k = \left(\frac{1}{1/-\theta s}\right)^k. \tag{2.41}$$

By comparing (2.40) with (2.41) we see that the sum of N exponentials becomes a Gamma distributed random variable with scale parameter $1/\lambda$ and shape parameter N. ■

While all the three generating functions are valid for a continuous random variable, only the probability-generating function is used for discrete RVs. Generalization to vector-valued random variables is straightforward. For a vector-valued random variable X the characteristic function is

$$\Phi_X(\boldsymbol{\omega}) = E_X[e^{\mathrm{i}\boldsymbol{\omega}^\mathsf{T}\mathbf{x}}] = \int_\infty p(\mathbf{x})e^{\mathrm{i}\boldsymbol{\omega}^\mathsf{T}\mathbf{x}}\mathrm{d}\mathbf{x}. \tag{2.42}$$

While we now have a tool for finding the distributions of sums of random variables, we lack a systematic methodology for discovering the distributions that result when a scalar or vector-valued random variable is subject to a non-linear transformation. While closed-form solutions in most such cases are unattainable, fundamental formulas nevertheless exist.

Theorem 2.4.1 — Nonlinear transformations of random variables. Suppose that $\mathbf{y} = \mathbf{f}(\mathbf{x})$ where $\mathbf{h} : \mathbb{R}^m \to \mathbb{R}^n$. Denote the pdf's of $\mathbf{x}$ and $\mathbf{y}$ by $g(\mathbf{x})$ and $h(\mathbf{y})$, respectively. Then we have that

$$h(\mathbf{y}) = \sum_i g(\mathbf{f}_i^{-1}(\mathbf{y}))\,|\det(\mathbf{F}_i^{-1}(\mathbf{y}))|$$

where $\mathbf{f}_i^{-1}(\mathbf{y})$ range over all solutions of $\mathbf{y} = \mathbf{f}(\mathbf{x})$ with respect to $\mathbf{x}$, and $\mathbf{F}_i^{-1}(\mathbf{y})$ is the corresponding Jacobian matrix of the inverse mapping $\mathbf{f}_i^{-1}(\mathbf{y})$.

Proof. For a proof in the 2-dimensional case, see [60] pages 199-201. ■

The sum in Theorem 2.4.1 reduces to a single term whenever $\mathbf{f}$ is invertible.

■ **Example 2.10 — Amplitude and phase of circularly symmetric Gaussian.** Let $\mathbf{x} = [x_1, x_2]^\mathsf{T}$ be a 2-dimensional random variable given by

$$g(\mathbf{x}) = \mathcal{N}(\mathbf{x}_1\,; 0, \sigma^2)\mathcal{N}(\mathbf{x}_2\,; 0, \sigma^2) = \mathcal{N}(\mathbf{x}; \mathbf{0}, \sigma^2\mathbf{I}). \tag{2.43}$$

In the last expression we have introduced the notation of the multivariate Gaussian, which will be extensively studied in Chapter 3. We want to find the pdf of the 2-dimensional random variable $\mathbf{y}$ given by $\mathbf{y} = \mathbf{f}(\mathbf{x})$ where $\mathbf{f} : \mathbb{R}^2 \to [0,\infty) \times [0, 2\pi)$ is given by

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} \|\mathbf{x}\|_2 \\ \text{atan2}(x_2, x_1) \end{bmatrix}. \tag{2.44}$$

Solution: First we notice that $\mathbf{f}$ is invertible on the given domain, since it simply is a conversion from Cartesian to polar coordinates. We can therefore drop the subscript i. Let us denote the components of $\mathbf{y}$ by r and θ, so that $\mathbf{y} = [r, \theta]^\mathsf{T}$. The inverse mapping and its Jacobian are given by

$$\mathbf{f}^{-1}(\mathbf{y}) = \begin{bmatrix} r\cos\theta \\ r\sin\theta \end{bmatrix} \quad \text{and} \quad \mathbf{F}^{-1}(\mathbf{y}) = \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix}. \tag{2.45}$$

It is easy to see that $|\mathbf{F}^{-1}(\mathbf{y})| = r$, and it follows that

$$h(\mathbf{y}) = \frac{r}{2\pi\sigma^2}\exp\left(-\frac{1}{2\sigma^2}\left(r^2\cos^2\theta + r^2\sin^2\theta\right)\right) = \frac{r}{2\pi\sigma^2}\exp\left(-\frac{r^2}{2\sigma^2}\right) \tag{2.46}$$

While the expression in (2.46) technically solves the given problem, we are not entirely satisfied before we have the marginal pdfs of r and θ. Since only r is present in (2.46), and since θ is defined on an interval of length 2π, we do not need to perform any aditional calculations to find these: We simply factorize the joint density as $h(\mathbf{y}) = p_r(r)p_\theta(\theta)$ where

$$p(r) = \frac{r}{\sigma^2}\exp\left(-\frac{r^2}{2\sigma^2}\right) = \text{Rayleigh}(r\,; \sigma^2) \tag{2.47}$$

$$p(\theta) = \frac{1}{2\pi}\chi_{[0,2\pi]}(\theta) = \text{Uniform}(\theta\,; [0, 2\pi]). \tag{2.48}$$

■

■ **Example 2.11 — Square of zero-mean univariate Gaussian.** If $X \sim \mathcal{N}(0,1)$, what is then the pdf of $Y = X^2$?

Solution: In this case, the mapping $\mathbf{f}$ in Theorem 2.4.1 is not invertible. For any realization of Y, we have two possible realizations of X:

$$\mathbf{f}_1^{-1} = \sqrt{y} \quad \text{and} \quad \mathbf{f}_2^{-1} = -\sqrt{y}. \tag{2.49}$$

The corresponding Jacobians are

$$\mathbf{F}_1^{-1}(y) = \frac{1}{2\sqrt{y}} \quad \text{and} \quad \mathbf{F}_2^{-1}(y) = \frac{-1}{2\sqrt{y}}. \tag{2.50}$$

By stitching this together, we arrive at

$$h(y) = \frac{1}{\sqrt{2\pi}}\left[\frac{1}{2\sqrt{y}}e^{-\frac{1}{2}(\sqrt{y})^2} + \frac{1}{2\sqrt{y}}e^{-\frac{1}{2}(-\sqrt{y})^2}\right] = \frac{1}{\sqrt{2\pi y}}e^{-y/2}. \tag{2.51}$$

We recognize this as a χ^2 distribution with 1 degree of freedom. ■

2.5 Frequentist and Bayesian approaches to probability

While everyone can agree on the fundamental framework outlined in the previous section, significant controversy exists when it comes to the more philosophical interpretation of what probability really is.

First, there is the question of whether randomness actually occur in nature, or whether randomness only serves as a model for stuff we do not have any better model for. In target tracking, the latter interpretation is much more relevant than the former, but this entails that we always must consider whether randomness actually provides a realistic model.

Second, there is a distinction between frequentist and Bayesian interpretation of probability. According to a die-hard frequentist, all probabilities should have an interpretation of a frequency of some sort. For example, if it rains in 250 out of 365 days in Bergen, then the probability of rain in Bergen is about 0.68. In contrast, most Bayesians are willing to consider statistical models that involve subjective assignments of probability. Such information is encoded in a prior distribution $p(X)$ which together with the likelihood $p(z|x)$ yields the posterior distribution

$$p(x|z) \propto p(z|x)p(x) \tag{2.52}$$

This allows the Bayesian to say that the unknown x has so and so probability for being this or that, or within a given interval. Frequentists will do no such thing. Instead a frequentist will analyze how plausible the data are given x. The philosophical difference between the two schools can be seen in the different interpretations of frequentist *confidence interval* and Bayesian *credible intervals*. A 95% confidence interval is constructed so that if the experiment is repeated several times, then it will cover the true and deterministic value of x 95% of the time. A 95% credible interval will contain the true and random value of x 95% of the time.

The capability of treating estimation in purely probabilistic terms is a huge advantage for the Bayesian approach. The Bayesian can always present the full posterior as a solution to his estimation problem, and ask the customer what information she would like to see extracted from it. The caveat is of course that the prior distribution, which necessarily has a subjective element, normally will play an important role in shaping the posterior. For some problems it may be appropriate to choose a so-called noninformative prior, which is designed to minimize the amount of information imposed by the prior. In other cases, physical models do indeed provide prior knowledge that it would be rather silly not to utilize. This is typically the case in sensor fusion.

2.5.1 Bayes and conditional probability for continuous RVs

While Bayes' rule itself is not a controversial result, it is nevertheless not obvious for continuous random variables, but something that must be proved.

Theorem 2.5.1 — Bayes' rule for pdfs. Let the continuous random variables X and Z have the joint distribution $p(x,z)$. Then

$$p(x|z) = \frac{p(z|x)p(x)}{p(z)} \tag{2.53}$$

Proof. For simplicity assume that both X and Z are scalar. Let us first define the events

$$\begin{aligned} A &= \{\, z \le Z \le z + \Delta z \,\}. \\ B &= \{\, x \le X \le x + \Delta x \,\}. \end{aligned} \tag{2.54}$$

Let us denote all the cdfs of x and z by $P(\cdot)$, with and without conditioning on A or B. Bayes' law

for probabilities yields

$$\frac{\Pr\{A\,|\,B\}}{\Pr\{A\}} = \frac{\Pr\{B\,|\,A\}}{\Pr\{B\}} \tag{2.55}$$

We can rewrite this expression in terms of cdfs as follows.

$$\frac{\dfrac{P(z+\Delta z\,|\,B)-P(z\,|\,B)}{\Delta z}}{\dfrac{P(z+\Delta z)-P(z)}{\Delta z}} = \frac{\dfrac{P(x+\Delta x\,|\,A)-P(x\,|\,A)}{\Delta x}}{\dfrac{P(x+\Delta x)-P(x)}{\Delta x}} \tag{2.56}$$

We have included division by Δz and Δx above and below the original fraction bar because we obviously intend to convert the cdfs to pdfs, which will involve differentiation. Let us then define the conditional pdfs according to

$$\begin{aligned} p(z\,|\,x) &= \lim_{\Delta x\to 0} p(z\,|\,B) = \lim_{\Delta z\to 0}\lim_{\Delta x\to 0} \frac{P(z+\Delta z\,|\,B)-P(z\,|\,B)}{\Delta z} \\ p(x\,|\,z) &= \lim_{\Delta z\to 0} p(x\,|\,A) = \lim_{\Delta x\to 0}\lim_{\Delta z\to 0} \frac{P(x+\Delta x\,|\,A)-P(x\,|\,A)}{\Delta x}. \end{aligned} \tag{2.57}$$

If we now take the double limit of (2.56) for $\Delta x \to 0$ and $\Delta z \to 0$ we can insert these pdfs in the numerators, while the denominators obviously turn into $p(z)$ and $p(x)$. Thus, (2.56) becomes

$$\frac{p(z\,|\,x)}{p(z)} = \frac{p(x\,|\,z)}{p(x)} \tag{2.58}$$

which is nothing more than a restatement of Bayes' rule for pdfs. ■

It is important to keep in mind that all results concerning pdf's of continuous random variables, including Bayes and conditional probability, are based on differentiation of proper probabilities. Differentiation is always relative to the metric properties, i.e., parametrization and units, of the underlying space, which may be $\mathbb{R}$, $\mathbb{R}^d$ or something more exotic, e.g., a sphere. The choice of coordinate system may not only affect the values of the pdf's, but also their overall shape.

2.6 Estimators

Estimation is the task of inferring knowledge about an unknown quantity x from data z which are related to x. In the probabilistic paradigm, the relationship between z and x is in the form of a probabilistic model $p(z\,|\,x)$. In the frequentist approach, this is all that is given. The Bayesian approach, prior knowledge about x is also given in the form of another probabilistic model $p(x)$, and one must then weigh the two models against each other to find an estimate of x.

For a given estimation problem, an *estimator* is a procedure that attempts to guess a concrete value for what x is based on the data. The output of the estimator is called an *estimate*.

Definition 2.6.1 — Estimator. Let $\mathscr{Z}$ and $\mathscr{X}$ be the spaces in which z and x belong, respectively. An estimator is a function $\theta : \mathscr{Z} \to \mathscr{X}$ so that $\theta(z)$ gives an estimate of x. We use a hat to denote the estimate, i.e., $\hat{x} = \theta(z)$.

Estimators can be categorized according to various categories. First, there is of course the divide between frequentist and Bayesian estimators. We may also distinguish between *maximizing* and *averaging* estimators. A maximizing estimator aims to find the one best value of x given z. An averaging estimator aims to find a value of x which is representative of our knowledge about x given z. While maximizing estimators can be defined in both frequentist and Bayesian estimation, the concept of an averaging estimator necessarily involves elements of Bayesian thinking. If one wants to have such tools available, it is hard remain a die-hard frequentist.[1]

[1] "Everyone is either a Bayesian or a closet Bayesian", Josef Knoll.

2.6.1 Maximizing estimators

Two well-known maximizing estimators are the maximum likelihood (ML) and maximum *a posteriori* (MAP) estimators. The ML estimator is given by

$$\hat{x} = \arg\max_x p_{Z|X}(z|x). \tag{2.59}$$

while the MAP estimator is given by

$$\hat{x} = \arg\max_x p_{X|Z}(x|z) = \arg\max_x p_{Z|X}(z|x)p_X(x). \tag{2.60}$$

In some special cases closed-form expressions for the ML or MAP estimators exist. It does probably not come as a big surprise that such cases include estimation of expectation and covariance for a Gaussian distribution. Another example is given in Example 2.12. In general, however, numerical search techniques are needed to implement these estimators. The menu that such techniques can be chosen from is fairly daunting, and a solid understanding of the objective functions in (2.59) or (2.60) is essential, both to succeed at all, and to achieve acceptable computational efficiency.

To which extent are the ML and MAP estimators reliable? The answer is to a larger extent affirmative for the ML estimator than for the MAP estimator. The ML estimator has some nice properties that statisticians refer to as *consistency* and *efficiency*. The former means that given enough data it will always recover the correct parameter. The latter means that it as it approaches that limit will attain a higher accuracy than any other kind of estimator that utilizes the same information.

For the MAP estimator we need to be more careful, since it also involves a prior distribution and Bayes' rule. Recalling that pdf's are only defined by differentiation relative to the coordinate system, it is clear that a change of coordinates can alter the location of the peak of a pdf. Since the MAP estimator maximizes the posterior pdf, it is therefore clear that the MAP estimator may not be invariant under coordinate transformations. This should not, however, be taken as a criticism of the Bayesian paradigm. A bona-fide Bayesian will never think of the MAP estimator as the solution to her estimation problem. In the Bayesian mindset, the entire posterior distribution is the solution, and more or less reliable information about this solution can be extracted by various estimators. For discrete estimation problems, also known as classification problems, the picture is somewhat different. Then there is a (hopefully nonzero) probability that the estimator hits the right value or class, and it can be shown that the MAP estimator has the highest success rate of all possible estimators, see Exercise 2.7.

2.6.2 Estimators as random variables

Since an estimator depends on the data, and the data both in Bayesian and frequentist schools are random variables, estimators are random variables as well. This means that a given estimator for a given model has a particular distribution. It is important to be aware of this for several reasons. Often, one wants to know the mean square error (MSE) of an estimator, and the MSE is given by this distribution. Furthermore, knowledge about an estimator's distribution can be important in subsequent processing of the estimate. For example, if two different estimates are to be combined in a sensor fusion system, then we would typically put most weight on the estimate with the lowest MSE.

■ **Example 2.12 — ML estimator of Rayleigh distribution.** Let $\mathbf{z} = [z_1, \ldots, z_M]^\mathsf{T}$ consist of i.i.d. samples from a Rayleigh distribution:

$$p(\mathbf{z}\,|\,\eta) = \prod_{i=1}^{M} \frac{z_i}{\eta} \exp\left(-\frac{z_i^2}{\eta^2}\right). \tag{2.61}$$

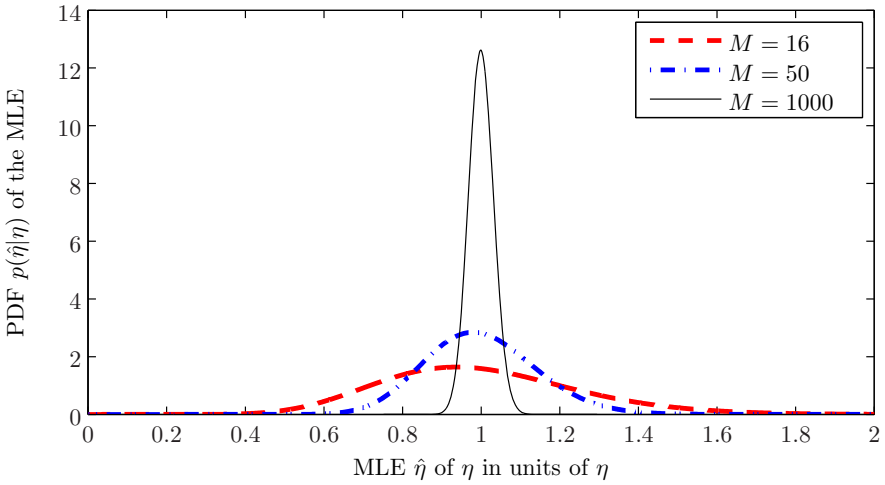


Figure 2.5: Distribution of the Rayleigh MLE for different sample sizes. Notice that a rather large number of samples is needed if η is to be estimated with accuracy of say 5%.

By differentiating the logarithm of $p(\mathbf{z}\,|\,\eta)$ and equating the derivative to zero, we get the ML estimator

$$\hat{\eta} = \frac{1}{2M}\sum_{i=1}^{M} z_i^2. \tag{2.62}$$

It can be shown that this quantity has a Gamma distribution. In Exercise 2.3 you are asked to show that each z_j^2 has an exponential distribution. Furthermore, we know from Example 2.9 that such a sum of exponential random variables is Gamma distributed. Through these steps we arrive at

$$p(\hat{\eta}\,|\,\eta) = \text{Gamma}\left(\hat{\eta}\,; M, \frac{2\eta}{M}\right) = \frac{M^M}{\Gamma(M)}\left(\frac{\hat{\eta}}{\eta}\right)^M \exp\left(-M\frac{\hat{\eta}}{\eta}\right). \tag{2.63}$$

This result is of central importance in detection theory, because we can use it to tune detection probabilities and false alarm rates of radar detectors. ∎

Related to the concept of estimators is the concept of *statistics*. A statistic is a single measure of some attribute of a sample. Estimators are statistics, but we can also construct statistics that not necessarily correspond to a meaningful estimator. As an example, the term $\sum_{i=1}^{M} z_i^2$ in Example 2.12 is a statistic. When we talk about statistics in this sense, we are primarily interested in *sufficient statistics*.

Definition 2.6.2 — Sufficient statistic. A sufficient statistic $g(\mathbf{z})$ for the parameter $\mathbf{x}$ summarizes the information about $\mathbf{x}$ contained in the data $\mathbf{z}$.

Returning to Example 2.12, we see that the term $\sum_{i=1}^{M} z_i^2$ is also a sufficient statistic for the parameter η. This is evident, because in (2.62) the individual data values z_j only occur in this expression. In other words, there is no dependence of the data which are not encapsulated by this term, and the term is then by definition a sufficient statistic. More generally, according to the Neyman-Fisher factorization theorem, it holds that $g(\mathbf{z})$ is a sufficient statistic if the likelihood can be factorized as $f(\mathbf{z}\,|\,\mathbf{x}) = f_1(g(\mathbf{z}),\mathbf{x})f_2(\mathbf{z})$. By a slight abuse of terminology, the parameters required to describe a given distribution, e.g., expectation and covariance of a Gaussian, are sometimes referred to as sufficient statistic. This is abusive because these are not functions of the data. On the other hand, the sample mean and the sample covariance *are* statistics.

2.6.3 LS and MMSE estimators

The least squares (LS) estimator and the minimum mean square error (MMSE) estimator both represent philosophies different from the the ML or MAP estimators. Consider an estimation problem of the form $\mathbf{z} = \mathbf{h}(\mathbf{x}) + \mathbf{w}$ where $\mathbf{w}$ is an unknown vector of measurement noise. The least squares estimator is

$$\hat{\mathbf{x}}_{\text{LS}} = \arg\min_{\mathbf{x}} \|\mathbf{z} - \mathbf{h}(\mathbf{x})\|_2^2. \tag{2.64}$$

It is evident from (2.64) that the LS estimator does not make any assumptions about the measurement "noise" $\mathbf{w}$. It is therefore of a non-probabilistic nature, and is a prime candidate to be considered in problems where a probabilistic model is not available. If $\mathbf{w}$ consists of i.i.d. Gaussian samples, then the LS estimator becomes identical to the MLE.

The idea of minimizing quadratic cost functions also plays a central role in the Bayesian paradigm. In accordance with the Bayesian philosophy of accounting for all uncertainty, the central concept in Bayesian decision theory is to make decisions that minimize given risk measures. An estimator that can be shown to minimize some risk function, is said to be Bayes-optimal with regard to that risk function. This is a very general framework. The decisions can in principle be any actions that are based on observing the data. The risk function can be any quantification of the consequences of the decision. Within the limited scope of estimation, however, the decisions will typically be nothing more than choosing a particular possibility of the estimatee as the estimate. In mathematical terms, if $l(\hat{\mathbf{x}}, \mathbf{x})$ is a loss function describing the deviation between the estimatee θ and our guess of its value $\hat{\theta}$, then the corresponding Bayes-optimal estimator is

$$\hat{\mathbf{x}}_l = \arg\min_{\hat{\mathbf{x}}} E_{\mathbf{x}}[l(\hat{\mathbf{x}}, \mathbf{x})]. \tag{2.65}$$

The most popular risk function is the expectation of the squared error, which in the vector case means the expectation of the function

$$l(\hat{\mathbf{x}}, \mathbf{x}) = (\hat{\mathbf{x}} - \mathbf{x})^{\mathsf{T}}(\hat{\mathbf{x}} - \mathbf{x}). \tag{2.66}$$

The estimator that results is known as the minimum mean square error (MMSE) estimator, which also is known as the expected *a posteriori* estimator, for reasons that will become obvious in the following theorem.

Theorem 2.6.1 — MMSE estimator. The MMSE estimator is given by

$$\hat{\mathbf{x}}_{\text{MMSE}} = \arg\min_{\hat{\mathbf{x}}} E\left[(\hat{\mathbf{x}} - \mathbf{x})^{\mathsf{T}}(\hat{\mathbf{x}} - \mathbf{x})\right] = E[\mathbf{x}|\mathbf{z}] = \int \mathbf{x} p(\mathbf{x}|\mathbf{z}) \mathrm{d}\mathbf{x}. \tag{2.67}$$

Proof. If $\hat{\mathbf{x}}_{\text{MMSE}}$ minimizes $E[l(\hat{\mathbf{x}}, \mathbf{x}) \,|\, \mathbf{z}]$ for all $\mathbf{z}$, then it will also minimize $E[l(\hat{\mathbf{x}}, \mathbf{x})]$. We proceed to show that it indeed does that by means of differentiation.

$$\frac{\partial E[(\hat{\mathbf{x}} - \mathbf{x})^{\mathsf{T}}(\hat{\mathbf{x}} - \mathbf{x}) \,|\, \mathbf{z}]}{\partial \hat{\mathbf{x}}} = E\left[\frac{\partial (\hat{\mathbf{x}} - \mathbf{x})^{\mathsf{T}}(\hat{\mathbf{x}} - \mathbf{x})}{\partial \hat{\mathbf{x}}} \,|\, \mathbf{z}\right] = E\left[-2(\hat{\mathbf{x}} - \mathbf{x})^{\mathsf{T}} \,|\, \mathbf{z}\right] \tag{2.68}$$

To identify the minimizer of the MSE we equate this with zero, which yields

$$\mathbf{0} = E\left[\hat{\mathbf{x}} - \mathbf{x} \,|\, \mathbf{z}\right] = E\left[\hat{\mathbf{x}} \,|\, \mathbf{z}\right] - E\left[\mathbf{x} \,|\, \mathbf{z}\right]. \tag{2.69}$$

Since $\hat{\mathbf{x}}$ is a function of the data we have that $\hat{\mathbf{x}} = E\left[\hat{\mathbf{x}} \,|\, \mathbf{z}\right]$ and it follows that $\hat{\mathbf{x}} = E\left[\mathbf{x} \,|\, \mathbf{z}\right]$. ■

2.6.4 Bias, MSE and variance of estimators

We generally want estimators that have low bias and MSE. Let $\tilde{\mathbf{x}} = \mathbf{x} - \hat{\mathbf{x}}$ be the estimation error. The bias of an estimator is then defined as the expectation $E[\tilde{\mathbf{x}}]$. We say that an estimator is *unbiased* if $E[\tilde{\mathbf{x}}] = 0$. The MMSE estimator is always unbiased. The ML and MAP estimators are, on the other hand, in general biased. Unbiasedness is obviously desirable in general. However, there may exist biased estimators with lower MSE than the best unbiased estimator, and there may exist estimation problems where the requirement of unbiasedness will lead to unacceptable degradation of the MSE.

In the scalar case, the variance and MSE of an estimator are given by

$$\mathrm{Var}(\hat{x}) = E[(\hat{x} - E[\hat{x}])^2] \ , \ \mathrm{MSE}(\hat{x}) = E[(\hat{x} - x)^2]. \tag{2.70}$$

From this it follows that the MSE can be decomposed into variance and bias as follows:

$$\mathrm{MSE}(\hat{x}) = \mathrm{Var}(\hat{x}) + \mathrm{Bias}(\hat{x}, x)^2. \tag{2.71}$$

In the vector case, the variance becomes a matrix:

$$\mathrm{Cov}(\hat{\mathbf{x}}) = E\left[(\hat{\mathbf{x}} - E[\hat{\mathbf{x}}])(\hat{\mathbf{x}} - E[\hat{\mathbf{x}}])^\mathsf{T}\right], \tag{2.72}$$

The MSE is on the other hand a scalar number also in the vector case, as already indicated by (2.67):

$$\mathrm{MSE}(\hat{\mathbf{x}}) = E\left[\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2\right] = E\left[(\hat{\mathbf{x}} - \mathbf{x})^\mathsf{T}(\hat{\mathbf{x}} - \mathbf{x})\right] = \mathrm{tr}\left(E[(\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^\mathsf{T}]\right). \tag{2.73}$$

2.6.5 LMMSE estimators

When the MMSE estimator is too complicated we may settle for the best linear estimator.

Theorem 2.6.2 — LMMSE estimator. The best estimator of the form $\hat{\mathbf{x}} = \mathbf{A}\mathbf{z} + \mathbf{b}$ that minimizes $\mathrm{MSE}(\hat{\mathbf{x}}) = E\left[\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2\right]$ is given by

$$\hat{\mathbf{x}} = E[\mathbf{x}] + \mathrm{Cov}(\mathbf{x}, \mathbf{z})\mathrm{Cov}(\mathbf{z})^{-1}(\mathbf{z} - E[\mathbf{z}]). \tag{2.74}$$

Proof. We prove the case where the estimatee is a scalar x, i.e., when the estimator is of the form $\hat{x} = \mathbf{a}^\mathsf{T}\mathbf{z} + b$. First we find what b should be. The estimation error and its square can be written

$$\tilde{x} = x - \mathbf{a}^\mathsf{T}\mathbf{z} - b \tag{2.75}$$

$$\tilde{x}^2 = (x - \mathbf{a}^\mathsf{T}\mathbf{z})^2 - 2b(x - \mathbf{a}^\mathsf{T}\mathbf{z}) + b^2, \tag{2.76}$$

respectively. The MSE is

$$E[\tilde{x}^2] = E[(x - \mathbf{a}^\mathsf{T}\mathbf{z})^2] - 2b(E[x] - \mathbf{a}^\mathsf{T}E[\mathbf{z}]) + b^2. \tag{2.77}$$

To find the value of b that minimizes this we calculate the derivative with respect to b and equate it with zero. This results in

$$b = E[x] - \mathbf{a}^\mathsf{T}E[\mathbf{z}]. \tag{2.78}$$

We could also have arrived at this result by requiring that LMMSE estimator should be unbiased. It is easy to see that if we insert (2.78) into the expression for $\tilde{x}$ in (2.75) then the bias becomes zero.

We proceed to look for the optimal vector $\mathbf{a}$. By demanding the derivative of the MSE with respect to $\mathbf{a}$ to be zero, we obtain

$$\frac{\partial}{\partial \mathbf{a}} E[\tilde{x}^2] = \frac{\partial}{\partial \mathbf{a}} E\left[(x - E[x] - \mathbf{a}(\mathbf{z} - E[\mathbf{z}]))^2\right] = 2E[\tilde{x}(\mathbf{z} - E[\mathbf{z}])^\mathsf{T}] = 0. \tag{2.79}$$

The second equality follows from the chain rule. The result that $E[\tilde{x}(\mathbf{z} - E[\mathbf{z}])^\mathsf{T}] = 0$ is known as the orthogonality principle. By further analyzing what this entails we obtain

$$\begin{aligned} E[\tilde{x}\mathbf{z}^\mathsf{T}] &= E\left[\left(x - E[x] - \mathbf{a}^\mathsf{T}(\mathbf{z} - E[\mathbf{z}]\right)(\mathbf{z} - E[\mathbf{z}])^\mathsf{T}\right] \\ &= \mathrm{Cov}[x, \mathbf{z}] - \mathbf{a}^\mathsf{T}\mathrm{Cov}[\mathbf{z}] = 0. \end{aligned} \tag{2.80}$$

This leads to the optimal choice of

$$\mathbf{a} = \mathrm{Cov}[x, \mathbf{z}]\,\mathrm{Cov}[\mathbf{z}]^{-1}. \tag{2.81}$$

Combining this with the expression for b yields the formula in the theorem. ■

It can also be seen that the matrix MSE of the LMMSE estimator is given by

$$\begin{aligned} E[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\mathsf{T}] =& E\Big[\left(\mathbf{x} - E[\mathbf{x}] - \mathrm{Cov}[\mathbf{x}, \mathbf{z}]\mathrm{Cov}[\mathbf{z}]^{-1}(\mathbf{z} - E[\mathbf{z}])\right) \\ & \qquad \left(\mathbf{x} - E[\mathbf{x}] - \mathrm{Cov}[\mathbf{x}, \mathbf{z}]\mathrm{Cov}[\mathbf{z}]^{-1}(\mathbf{z} - E[\mathbf{z}])\right)^\mathsf{T}\Big] \\ =& \mathrm{Cov}[\mathbf{x}] - \mathrm{Cov}[\mathbf{x}, \mathbf{z}]\mathrm{Cov}[\mathbf{z}]^{-1}\mathrm{Cov}[\mathbf{x}, \mathbf{z}]^\mathsf{T}. \end{aligned} \tag{2.82}$$

The LMMSE equations (2.74) and (2.82) bears some resemblance to the Kalman filter formulas. In fact, we could have derived the Kalman filter as an example of LMMSE estimation by means of the orthogonality principle. We shall instead, however, study the Kalman filter in a more directly Bayesian setting, because this will make it easier to understand how the Kalman filter is used in sensor fusion applications such as target tracking and SLAM. For this purpose, the next chapter is entirely devoted to the multivariate Gaussian distribution.

2.7 References and chapter remarks

The aim of this chapter has largely been to provide a condensed summary of the first three chapters in [2]. Thus, if the reader is looking for further details, that would be a good place to start. For fundamentals of probability theory, the reader may consult a basic probability textbook such as [79], or go straight to the more comprehensive [60]. The proof for Bayes' rule for pdf's in Theorem 2.5.1 is based on a somewhat more cursorial proof found in [60]. To do anything with the measure theoretic approach to probability must students who follow TTK4250 would first need to become familiar with measure theory, which is rigorously explained in [26]. The best source on Bayesian estimation and decision theory known to the author is [24].

2.8 Exercises

Exercise 2.1 Let $Y = X_1 + \ldots + X_n$ be the sum of n independent random vectors. Show that the characteristic function of Y as given by (2.42) is equal to the product of the characteristic functions of $X_1, \ldots, X_n$. ■

Exercise 2.2 Let $X_1, \ldots, X_n$ be n i.i.d. random variables, each χ^2 distributed with k degrees of freedom. Show that $Y = X_1 + \ldots + X_n$ is χ^2 distributed with nk degrees of freedom. ■

Exercise 2.3 Let X be a Rayleigh distributed random variable with parameter σ^2 as defined on page 21. Show that X^2 is an exponential random variable with parameter $\lambda = 1/\sigma^2$. ■

Exercise 2.4 Prove the Lindeberg-Lévy version of the CLT. That is, prove Theorem 2.3.1. **Hint:** Use generating functions. ■

Exercise 2.5 Let $X_1, \ldots, X_n$ be n independent Bernoulli random variables with success probabilities $r_1, \ldots, r_n$, respectively.

a) What is distribution of $Y = \sum_{i=1}^{n} X_n$?
b) What is the expectation of Y?
c) What is the covariance of Y?

■

Exercise 2.6 Let X be a random variable with cumulative distribution function F. Show that the random variable $Y = F(X)$ is uniformly distributed over $[0,1]$. ■

Exercise 2.7 Show that the MAP classifier has the highest success rate of all possible classifiers for a Bayesian classification problem. ■

3. The multivariate Gaussian

The key construction that underlies the Kalman filter, and virtually all of sensor fusion, is the multivariate Gaussian, which generalizes the univariate Gaussian from Example (2.6). The distribution of a multivariate Gaussian RV is given by its expectation vector μ and symmetric positive definite covariance matrix $\mathbf{P}$, according to

$$\mathcal{N}(\mathbf{x};\mu,\mathbf{P}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{P}|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^{\mathsf{T}}\mathbf{P}^{-1}(\mathbf{x}-\mu)\right) \tag{3.1}$$

Recall that the univariate Gaussian pdf always looks like a bell curve whose peak location and spread are governed by its expectation and covariance, respectively. A two-dimensional Gaussian gives a similar bell surface, whose peak location is given by its expectation vector, and whose shape is given by its covariance matrix. This generalizes in the obvious manner to higher dimensions.

3.1 Quadratic forms and covariance ellipses

The exponent in (3.1) is a quadratic form in the variable $\mathbf{x}$. Let us more precisely define the quadratic form corresponding to (3.1) as

$$q(\mathbf{x}) = (\mathbf{x}-\mu)^{\mathsf{T}}\mathbf{P}^{-1}(\mathbf{x}-\mu). \tag{3.2}$$

The value of this function for a given $\mathbf{x}$ and μ is also known as the Mahalanobis distance between $\mathbf{x}$ and μ. In terms of this function we can write the logarithm of the Gaussian pdf as

$$\ln\mathcal{N}(\mathbf{x};\mu,\mathbf{P}) = c - \frac{1}{2}q(\mathbf{x}) \tag{3.3}$$

where c is a constant that is given by the normalization requirement, and therefore carries no information about $\mathbf{x}$. From this we make three very important observations.

First, this means that the curvature, or more precisely the Hessian, of $\ln\mathcal{N}(\mathbf{x};\mu,\mathbf{P})$ is constant. In fact, if $\mathcal{H}_{\mathbf{x}}$ denotes the Hessian, the following identity holds for any $\mathbf{x}$:

$$\mathcal{H}_{\mathbf{x}}\ln\mathcal{N}(\mathbf{x};\mu,\mathbf{P}) = -\mathbf{P}^{-1}. \tag{3.4}$$

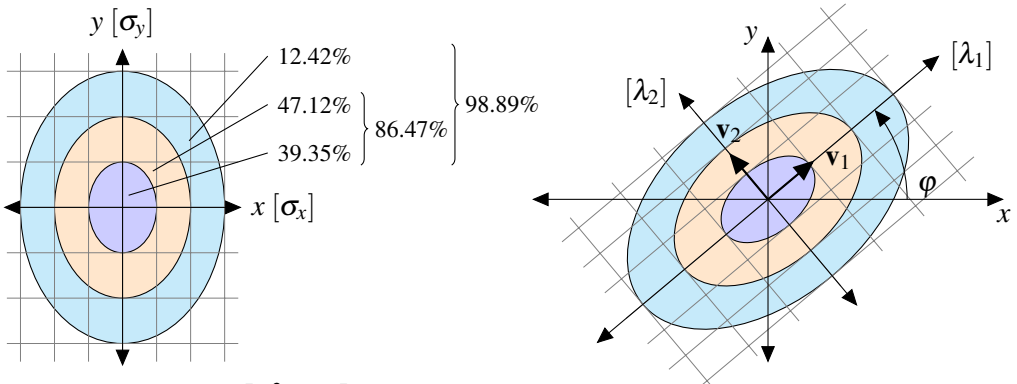
Second, the level curves for a 2-dimensional Gaussian are ellipses, since any expression of the form $Ax^2 + Bxy + Cy^2 = 1$ describes an ellipse in the $x-y$-plane. More generally, we get ellipsoids for a 3-dimensional Gaussian, and so on. For simplicity, we will just refer to these as covariance ellipses, irrespectively of the dimension. The axes of these ellipses can be found by means of eigenvalue decomposition of $\mathbf{P}$. If $(\lambda_i, \mathbf{e}_i)$ are the eigenvalues and eigenvectors of $\mathbf{P}$, the the 1σ ellipse of $\mathbf{P}$ has axes given by $\sqrt{\lambda_i}\mathbf{e}_i$, and the more general $g\sigma$-ellipse has axes given by $g\sqrt{\lambda_i}\mathbf{e}_i$. The area, or more generally hypervolume, within such an ellipse is given by

$$\frac{\pi^{n/2}}{\Gamma(n/2+1)} g^n \sqrt{|\mathbf{P}|}.$$

Third, it is evident that a Gaussian is entirely specified in terms of its quadratic form. Conversely, if we multiply a convex quadratic form in $\mathbf{x}$ by $\left(-\frac{1}{2}\right)$, and then exponentiate it, we get a function which is proportional to a particular Gaussian distribution in $\mathbf{x}$. The normalization requirement implies that this function cannot be proportional to any other pdf than this particular Gaussian. The consequence of this is that when we study how multivariate Gaussians behave under various operations (marginalization, conditioning, etc.) *we only need to study the quadratic forms*.

■ **Example 3.1 — Conditioning is entirely given by quadratic forms.** Let $p(x,y)$ be a bivariate Gaussian, and let $f(x,y)$ be its corresponding quadratic form. The conditional distribution of x given y is $p(x|y) = p(x,y)/p(y)$. We know that both $p(x|y)$, $p(x,y)$ and $p(y)$ are valid pdfs. We also know that $p(y)$ is a function of y alone. The dependency of $p(x|y)$ on x is therefore given by the quadratic form $f(x,y)$, this time treated as a function of x alone, with y fixed. Thus, we see that $p(x|y)$ also must be Gaussian, and that we can determine $p(x|y)$ from $p(x,y)$ just by studying quadratic forms. The exact details for how this is to be done are left for Theorem 3.2.3 in the next section. ■

The amount of probability within a $g\sigma$-ellipse decreases as the dimension n increases. This is because higher dimension provides more directions in which probability mass can be spread. to quantify this, the trick is to transform the Gaussian random vector into a scalar χ^2 random variable whose cdf can be written in terms of the error function or directly evaluated using, e.g., Matlab. More precisely, if $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{P})$ it can be shown that $(\mathbf{x}-\mu)^{\mathsf{T}}\mathbf{P}^{-1}(\mathbf{x}-\mu)$ has a χ^2 distribution with n degrees of freedom (Exercise 3.4). The probability mass within 1, 2 and 3 σ for a bivariate Gaussian is illustrated in Figure 3.1.



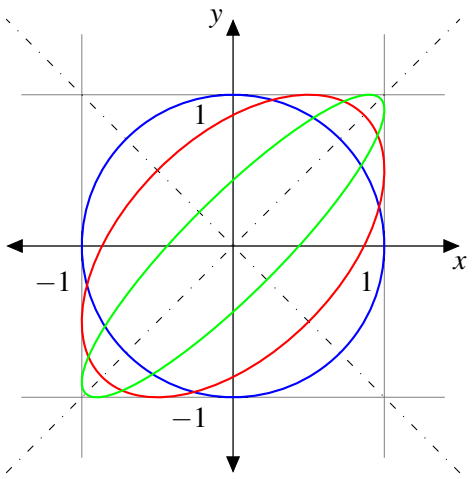
(a) $\Sigma = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix}$. (b) $\Sigma = \mathbf{P}\Lambda\mathbf{P}^T$, $\mathbf{P} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 \end{bmatrix}$

Figure 3.1: Two-dimensional Gaussian distribution. Probability of ellipses.

In the right hand side of Figure 3.1 we see how the similarity transform (also known as the eigendecomposition or spectral transform) of the covariance matrix yields an alternative coordinate

system, spanned by the eigenvectors, where all cross-correlations vanish. This can be useful for a variety of computational purposes. It is also possible to transform a correlated Gaussian vector to a correlation-free Gaussian vector by means of other decompositions such as the Cholesky factorization (see Example 3.4). We see that correlations make the covariance ellipses tilted. In the right hand side of Figure 3.1 there is a positive correlation between x and y. The ellipses would have been tilted the other way if it was negative.

■ **Example 3.2 — Correlations make the covariance ellipses narrower.** In Figure 3.2 we see the 1σ covariance ellipses of three bivariate Gaussians with unity marginal variance in the x- and y-directions. The top and bottom of all the covariance ellipses touch the -1 and $+1$ lines, and the same happens horizontally. However, the Gaussians with the highest cross-correlation a, have the narrowest covariance ellipses. This shows that the presence of correlations actually decrease uncertainty. For this reason, estimation methods such as the Kalman filter exploit correlations.



$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix} \right)$$

$$\tfrac{1}{2} \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix}^{-1} \begin{bmatrix} x \\ y \end{bmatrix} = 1$$

$a = 0.0$
$a = 0.5$
$a = 0.9$

Figure 3.2: For $a \in (-1,1)$, the semiaxes are $\sqrt{1+a}$ and $\sqrt{1-a}$ long and lie on the $y = x$ and $y = -x$ lines, respectively.

■

3.2 Rules for working with Gaussians

We will now establish several important results that make it much easier to work with with multivariate Gaussians than with any other kind of multivariate RVs. Based on the argument elaborated in Example 3.1, we will prove all these results by simply studying how the quadratic form in the exponent of (3.1) behaves.

Theorem 3.2.1 — Independence. Two random vectors $\mathbf{x}$ and $\mathbf{y}$ with probability density functions $\mathcal{N}(\mathbf{x};\mathbf{a},\mathbf{A})$ and $\mathcal{N}(\mathbf{y};\mathbf{b},\mathbf{B})$ are independent if and only if

$$p(\mathbf{x},\mathbf{y}) = p\left(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}; \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & 0 \\ 0 & \mathbf{B} \end{bmatrix}\right). \tag{3.5}$$

Proof. We need to prove that zero covariance implies that $p(\mathbf{x},\mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$ and vice versa. According to the inversion rule of Appendix Q.1 we know that the inverse of a block-diagonal matrix is found by simply inverting the blocks. Thus, the quadratic form in the exponent of the

joint Gaussian becomes

$$\begin{bmatrix} \mathbf{x}-\mathbf{a} \\ \mathbf{y}-\mathbf{b} \end{bmatrix}^\mathsf{T} \begin{bmatrix} \mathbf{A} & 0 \\ 0 & \mathbf{B} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{x}-\mathbf{a} \\ \mathbf{y}-\mathbf{b} \end{bmatrix} = \begin{bmatrix} (\mathbf{x}-\mathbf{a})^\mathsf{T} & (\mathbf{y}-\mathbf{b})^\mathsf{T} \end{bmatrix} \begin{bmatrix} \mathbf{A}^{-1} & 0 \\ 0 & \mathbf{B}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{x}-\mathbf{a} \\ \mathbf{y}-\mathbf{b} \end{bmatrix}$$
$$= (\mathbf{x}-\mathbf{a})^\mathsf{T}\mathbf{A}^{-1}(\mathbf{x}-\mathbf{a}) + (\mathbf{y}-\mathbf{b})^\mathsf{T}\mathbf{A}^{-1}(\mathbf{y}-\mathbf{b})$$

which is a sum of the quadratic forms from the marginal distributions. Based on this, we see that zero covariance implies that $\ln p(\mathbf{x},\mathbf{y}) = \ln p(\mathbf{x}) + \ln p(\mathbf{y})$, which again implies that $p(\mathbf{x},\mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$. The converse implications also follow. ■

This relationship between independence and zero covariance does not hold for arbitrary non-Gaussian random vectors.

There is a strong relationship between Gaussianity and linearity, to the extent that these concepts often are used interchangeably in the literature. If a Gaussian RV goes through a linear transform, the new RV is also Gaussian.

Theorem 3.2.2 — Linearity. If $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{a}, \mathbf{A})$ and $\mathbf{y} = \mathbf{F}\mathbf{x}$, then $p(\mathbf{y}) = \mathcal{N}(\mathbf{y}; \mathbf{F}\mathbf{a}, \mathbf{F}\mathbf{A}\mathbf{F}^\mathsf{T})$.

Proof. A general proof of this property can be constructed by means of the moment-generating function. In the special case that $\mathbf{F}$ is square and invertible, a simpler proof can be constructed by means of the nonlinear transformation formula in Theorem 2.4.1. We only sketch this proof here. Denote the original pdf by $g(\mathbf{x})$ and the pdf of the new RV by $h(\mathbf{y})$. We then have that

$$\begin{aligned} g(\mathbf{x}) &\propto \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{a})^\mathsf{T}\mathbf{A}^{-1}(\mathbf{x}-\mathbf{a})\right) \\ h(\mathbf{y}) &\propto \exp\left(-\frac{1}{2}(\mathbf{F}^{-1}\mathbf{y}-\mathbf{a})^\mathsf{T}\mathbf{A}^{-1}(\mathbf{F}^{-1}\mathbf{y}-\mathbf{a})\right) \\ &= \exp\left(-\frac{1}{2}(\mathbf{F}^{-1}\mathbf{y}-\mathbf{a})^\mathsf{T}\mathbf{F}^\mathsf{T}(\mathbf{F}^\mathsf{T})^{-1}\mathbf{A}^{-1}\mathbf{F}^{-1}\mathbf{F}(\mathbf{F}^{-1}\mathbf{y}-\mathbf{a})\right) \\ &= \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{F}\mathbf{a})^\mathsf{T}(\mathbf{F}^\mathsf{T})^{-1}\mathbf{A}^{-1}\mathbf{F}^{-1}(\mathbf{y}-\mathbf{F}\mathbf{a})\right). \end{aligned} \tag{3.6}$$

The theorem then follows from recognizing that $(\mathbf{F}^\mathsf{T})^{-1}\mathbf{A}^{-1}\mathbf{F}^{-1} = (\mathbf{F}\mathbf{A}\mathbf{F}^\mathsf{T})^{-1}$. ■

A related result is that if $\mathbf{x}$ has the distribution $\mathcal{N}(\mathbf{a},\mathbf{A})$, then the random vector $\mathbf{y} = \mathbf{x}+\mathbf{b}$ has the distribution $\mathcal{N}(\mathbf{a}+\mathbf{b},\mathbf{A})$. This follows almost trivially from the nonlinear transformation formula, because all we have to is shift $\mathbf{b}$ from the random vector slot to the expectation slot.

▪ **Example 3.3 — Sum of variances.** Let $x_1 \sim \mathcal{N}(0,\sigma^2)$ and $x_2 \sim \mathcal{N}(0,\sigma^2)$. Then it follows that $x = x_1 + x_2 \sim \mathcal{N}(0, 2\sigma^2)$. This follows from the theorem with $\mathbf{F} = [1,1]$, $\mathbf{x} = [x_1, x_2]^\mathsf{T}$ and $\mathbf{y} = x_3$. ▪

▪ **Example 3.4 — Cholesky decomposition of the covariance matrix.** Let $\mathbf{x} \sim \mathcal{N}(\mathbf{0},\mathbf{A})$. Since $\mathbf{A}$ is symmetric positive definite we know that it has a Cholesky factorization $\mathbf{L}$ so that $\mathbf{A} = \mathbf{L}\mathbf{L}^\mathsf{T}$. Let us then define the transformed RV $\mathbf{y} = \mathbf{L}^{-1}\mathbf{x}$. The expectation of $\mathbf{y}$ is then obviously zero as well. Its covariance becomes

$$\mathrm{Cov}[\mathbf{y}] = \mathbf{L}^{-1}\mathbf{A}(\mathbf{L}^{-1})^\mathsf{T} = \mathbf{L}^{-1}\mathbf{L}\mathbf{L}^\mathsf{T}(\mathbf{L}^{-1})^\mathsf{T} = \mathbf{I}. \tag{3.7}$$

The effect that the linear transformation $\mathbf{L}^{-1}$ has on $\mathbf{x}$ is often described as "whitening" or "prewhitening". This is a common and important operation, especially in signal processing, where $\mathbf{x}$ may be thought of as a long time sequence constituting a signal. By whitening a signal

we can perform operations on it (e.g, hypothesis tests) which require whiteness, i.e., independence between the samples. We can also use a reversal of the above argument to generate correlated Gaussian RVs. If we have a Gaussian RV $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then the transformed RV $\mathbf{x} = \mathbf{L}\mathbf{y}$ will have the covariance

$$\mathrm{Cov}[\mathbf{x}] = \mathbf{L}\mathbf{L}^{\mathsf{T}} = \mathbf{A}. \tag{3.8}$$

In Matlab, we can for example draw `N` indepedendent random vectors from $\mathcal{N}(\mathtt{a}, \mathtt{A})$ using

$$\mathtt{x} = \mathtt{repmat}(\mathtt{a}, [1, \mathtt{N}]) + \mathtt{cholA'*randn}(\mathtt{n}, \mathtt{N}). \tag{3.9}$$

The transpose is required because Matlab's `chol` function calculates the upper Cholesky matrix by default, and not the lower Cholesky matrix. ∎

Marginalization and conditioning are frequent tasks that must be done with the multivariate Gaussian. In the moment-based representation of the Gaussian, the former is trivial, while the latter is somewhat more complicated.

Theorem 3.2.3 — Marginalization and conditioning. Let $\mathbf{x}$ and $\mathbf{y}$ have the joint distribution

$$p(\mathbf{x}, \mathbf{y}) = \mathcal{N}\left(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}; \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathbf{P}_{xx} & \mathbf{P}_{xy} \\ \mathbf{P}_{xy}^{\mathsf{T}} & \mathbf{P}_{yy} \end{bmatrix}\right)$$

Then the marginal distribution of $\mathbf{y}$ is $p(\mathbf{y}) = \mathcal{N}(\mathbf{y}; \mathbf{b}, \mathbf{P}_{yy})$, and the conditional distribution pf $\mathbf{x}$ given $\mathbf{y}$ is $p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{x|y}, \mathbf{P}_{x|y})$ where $\boldsymbol{\mu}_{x|y} = \mathbf{a} + \mathbf{P}_{xy}\mathbf{P}_{yy}^{-1}(\mathbf{y} - \mathbf{b})$ and $\mathbf{P}_{x|y} = \mathbf{P}_{xx} - \mathbf{P}_{xy}\mathbf{P}_{yy}^{-1}\mathbf{P}_{xy}^{\mathsf{T}}$.

Proof. For simplicity, assume that $\mathbf{a}$ and $\mathbf{b}$ are zero. We use the matrix inversion rule (Q.3) to decompose the inverse covariance matrix as follows:

$$\begin{bmatrix} \mathbf{P}_{xx} & \mathbf{P}_{xy} \\ \mathbf{P}_{xy}^{\mathsf{T}} & \mathbf{P}_{yy} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{P}_{yy}^{-1}\mathbf{P}_{xy}^{\mathsf{T}} & \mathbf{I} \end{bmatrix} \begin{bmatrix} (\mathbf{P}_{xx} - \mathbf{P}_{xy}\mathbf{P}_{yy}^{-1}\mathbf{P}_{xy}^{\mathsf{T}})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{yy}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\mathbf{P}_{xy}\mathbf{P}_{yy}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}.$$

We insert this into the joint Gaussian, and obtain

$$p(\mathbf{x}, \mathbf{y}) \propto \exp\left(-\frac{1}{2}\begin{bmatrix} \mathbf{x} - \mathbf{P}_{xy}\mathbf{P}_{yy}^{-1}\mathbf{y} \\ \mathbf{y} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} (\mathbf{P}_{xx} - \mathbf{P}_{xy}\mathbf{P}_{yy}^{-1}\mathbf{P}_{xy}^{\mathsf{T}})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{yy}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{x} - \mathbf{P}_{xy}\mathbf{P}_{yy}^{-1}\mathbf{y} \\ \mathbf{y} \end{bmatrix}\right).$$

Since the center matrix is diagonal, this becomes

$$p(\mathbf{x}, \mathbf{y}) \propto \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{P}_{xy}\mathbf{P}_{yy}^{-1}\mathbf{y})^{\mathsf{T}}(\mathbf{P}_{xx} - \mathbf{P}_{xy}\mathbf{P}_{yy}^{-1}\mathbf{P}_{xy}^{\mathsf{T}})^{-1}(\mathbf{x} - \mathbf{P}_{xy}\mathbf{P}_{yy}^{-1}\mathbf{y})\right) \exp\left(-\frac{1}{2}\mathbf{y}^{\mathsf{T}}\mathbf{P}_{yy}^{-1}\mathbf{y}\right).$$

Both the desired results follow from inspection of this expression. We look at marginalization first. In this case, the goal is to find $p(\mathbf{y}) = \int p(\mathbf{x}, \mathbf{y})\mathrm{d}\mathbf{x}$. Irrespectively on the value of $\mathbf{y}$, the first exponential will describe a Gaussian in $\mathbf{x}$. For this reason, the integral over $\mathbf{x}$ eliminates this entire exponential, including its dependency on $\mathbf{y}$, and we are left with the second exponential. This proves the marginalization result. Then we look at conditioning. In this case, the goal is to find $p(\mathbf{x}|\mathbf{y}) = p(\mathbf{x}, \mathbf{y})/p(\mathbf{y})$. Carrying out this division amounts to removing the second exponential, so that we are left with only the first one. That is, the conditional density $p(\mathbf{x}|\mathbf{y})$ is a Gaussian with expectation $\mathbf{P}_{xy}\mathbf{P}_{yy}^{-1}\mathbf{y}$ and covariance $\mathbf{P}_{xx} - \mathbf{P}_{xy}\mathbf{P}_{yy}^{-1}\mathbf{P}_{xy}^{\mathsf{T}}$. Re-introducing the expectations $\mathbf{a}$ and $\mathbf{b}$ at their appropriate slots is straightforward, and concludes the proof. ∎

3.3 The product identity

After having tasted some samples of the niceness of the multivariate Gaussian in Sections 3.2 - 3.4, it should not come as any surprise that the product of two Gaussians is another Gaussian. After all, a Gaussian is what we get when we exponentiate a negative definite quadratic form, and when we add two quadratic forms we get another quadratic form, whose exponential becomes a Gaussian after normalization. What is perhaps more surprising is that the normalization constant also becomes a Gaussian.

Recall that the expression for a Gaussian, $\mathcal{N}(\mathbf{x};\mu,\mathbf{P})$, has two "slots" where vectors can be inserted, i.e., the argument slot and the expectation slot. What is particularly interesting is to see what happens when we have a product of two Gaussians, where $\mathbf{x}$ enters the argument slot in one of the Gaussians, and the expectation slot in the other Gaussian. This leads to the fundamental product identity.

Theorem 3.3.1 — The product identity. The following identity

$$\mathcal{N}(\mathbf{z};\mathbf{H}\mathbf{x},\mathbf{R})\mathcal{N}(\mathbf{x};\bar{\mathbf{x}},\bar{\mathbf{P}}) = \mathcal{N}(\mathbf{z};\bar{\mathbf{z}},\mathbf{S})\mathcal{N}(\mathbf{x};\hat{\mathbf{x}},\hat{\mathbf{P}}) \tag{3.10}$$

is true if the vectors and matrices involved are related according to

$$\begin{aligned}
\bar{\mathbf{z}} &= \mathbf{H}\bar{\mathbf{x}} \\
\hat{\mathbf{x}} &= \bar{\mathbf{x}} + \mathbf{W}(\mathbf{z} - \mathbf{H}\bar{\mathbf{x}}) \\
\mathbf{S} &= \mathbf{R} + \mathbf{H}\bar{\mathbf{P}}\mathbf{H}^{\mathsf{T}} \\
\hat{\mathbf{P}} &= (\mathbf{I} - \mathbf{W}\mathbf{H})\bar{\mathbf{P}} \\
\mathbf{W} &= \bar{\mathbf{P}}\mathbf{H}^{\mathsf{T}}\mathbf{S}^{-1}.
\end{aligned}$$

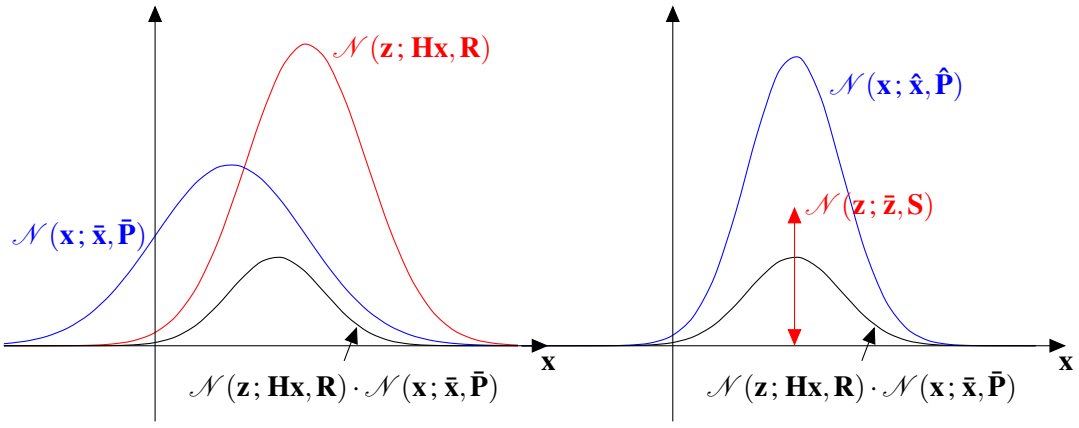


Figure 3.3: Illustration of the product identity.

Proof. We can prove the product identity by showing that both the left-hand-side and the right-hand-side are two possible factorizations of a joint Gaussian $p(\mathbf{x},\mathbf{z})$. In other words, we are going to use the relationships

$$p(\mathbf{x},\mathbf{z}) = p(\mathbf{z}|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z}). \tag{3.11}$$

Step 1: Construct the joint density. If we start by looking at the first factorization in (3.11), we may simply define $p(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{x})$ as identical to the two Gaussians on the left hand side of (3.10). This

also defines $p(\mathbf{x},\mathbf{z})$ as identical to their product, i.e.,

$$p(\mathbf{z},\mathbf{x}) = \mathcal{N}(\mathbf{z};\mathbf{H}\mathbf{x},\mathbf{R})\mathcal{N}(\mathbf{x};\bar{\mathbf{x}},\bar{\mathbf{P}}). \tag{3.12}$$

Step 2: Manipulate the quadratic form. The quadratic form in (3.12) can be written as follows:

$$\begin{aligned}
&(\mathbf{x}-\bar{\mathbf{x}})^{\mathsf{T}}\bar{\mathbf{P}}^{-1}(\mathbf{x}-\bar{\mathbf{x}}) + (\mathbf{z}-\mathbf{H}\mathbf{x})^{\mathsf{T}}\mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}\mathbf{x}) \\
&= \begin{bmatrix}\mathbf{x}-\bar{\mathbf{x}}\\ \mathbf{z}-\mathbf{H}\mathbf{x}\end{bmatrix}^{\mathsf{T}} \begin{bmatrix}\bar{\mathbf{P}}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}^{-1}\end{bmatrix} \begin{bmatrix}\mathbf{x}-\bar{\mathbf{x}}\\ \mathbf{z}-\mathbf{H}\mathbf{x}\end{bmatrix} \\
&= \begin{bmatrix}\mathbf{x}-\bar{\mathbf{x}}\\ \mathbf{z}-\mathbf{H}\bar{\mathbf{x}}\end{bmatrix}^{\mathsf{T}} \begin{bmatrix}\mathbf{I} & -\mathbf{H}^{\mathsf{T}}\\ \mathbf{0} & \mathbf{I}\end{bmatrix} \begin{bmatrix}\bar{\mathbf{P}}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}^{-1}\end{bmatrix} \begin{bmatrix}\mathbf{I} & \mathbf{0}\\ -\mathbf{H} & \mathbf{I}\end{bmatrix} \begin{bmatrix}\mathbf{x}-\bar{\mathbf{x}}\\ \mathbf{z}-\mathbf{H}\bar{\mathbf{x}}\end{bmatrix} \\
&= \begin{bmatrix}\mathbf{x}-\bar{\mathbf{x}}\\ \mathbf{z}-\mathbf{H}\bar{\mathbf{x}}\end{bmatrix}^{\mathsf{T}} \left(\begin{bmatrix}\mathbf{I} & \mathbf{H}\\ \mathbf{0} & \mathbf{I}\end{bmatrix} \begin{bmatrix}\bar{\mathbf{P}} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix} \begin{bmatrix}\mathbf{I} & \mathbf{H}^{\mathsf{T}}\\ \mathbf{0} & \mathbf{I}\end{bmatrix} \right)^{-1} \begin{bmatrix}\mathbf{x}-\bar{\mathbf{x}}\\ \mathbf{z}-\mathbf{H}\bar{\mathbf{x}}\end{bmatrix} \\
&= \begin{bmatrix}\mathbf{x}-\bar{\mathbf{x}}\\ \mathbf{z}-\mathbf{H}\bar{\mathbf{x}}\end{bmatrix}^{\mathsf{T}} \begin{bmatrix}\bar{\mathbf{P}} & \bar{\mathbf{P}}\mathbf{H}^{\mathsf{T}}\\ \mathbf{H}\bar{\mathbf{P}} & \mathbf{H}\bar{\mathbf{P}}\mathbf{H}^{\mathsf{T}}+\mathbf{R}\end{bmatrix}^{-1} \begin{bmatrix}\mathbf{x}-\bar{\mathbf{x}}\\ \mathbf{z}-\mathbf{H}\bar{\mathbf{x}}\end{bmatrix}
\end{aligned} \tag{3.13}$$

The second equality in this development is the only one that is not straightforward. The background for this equality is that we need to remove $\mathbf{x}$ from the outer vector if we are going to arrive at the product identity. This is done by subjecting the outer vector to an appropriate linear transform. More precisely, this is achieved by

$$\begin{bmatrix}\mathbf{I} & \mathbf{0}\\ -\mathbf{H} & \mathbf{I}\end{bmatrix} \begin{bmatrix}\mathbf{x}-\bar{\mathbf{x}}\\ \mathbf{z}-\mathbf{H}\bar{\mathbf{x}}\end{bmatrix} = \begin{bmatrix}\mathbf{x}-\bar{\mathbf{x}}\\ -\mathbf{H}\mathbf{x}+\mathbf{H}\bar{\mathbf{x}}+\mathbf{x}-\mathbf{H}\bar{\mathbf{x}}\end{bmatrix} = \begin{bmatrix}\mathbf{x}-\bar{\mathbf{x}}\\ \mathbf{z}-\mathbf{H}\mathbf{x}\end{bmatrix} \tag{3.14}$$

In other words, we can replace the right-hand-side of (3.14) with the left-hand-side of (3.14), which is done in the third line of (3.13).

Step 3: Factorize into marginal and conditional. The last expression in (3.13) is the quadratic form of a Gaussian in $\mathbf{x}$ and $\mathbf{z}$, and since it encapsulates all dependency on $\mathbf{x}$ and $\mathbf{z}$, we can rest assured that this Gaussian is the joint density $p(\mathbf{x},\mathbf{z})$. The covariance matrix in (3.13) is not block-diagonal, and therefore some further work is needed to split the quadratic form into a sum of two separate quadratic forms, as we need to arrive at the product identity. However, if we can construct the alternative factorization $p(\mathbf{x},\mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$ we may perhaps achieve this. To find the quadratic forms corresponding to $p(\mathbf{z})$ and $p(\mathbf{x}|\mathbf{z})$ we use Theorem 3.2.3, which provided expressions for marginal and conditional Gaussians. If we compare with the entities involved in Theorem 3.2.3, we can make the following identifications

Role	In Theorem 3.2.3	In (3.13)
Conditioned RV	$\mathbf{x}$	$\mathbf{x}$
RV that we condition on	$\mathbf{y}$	$\mathbf{z}$
Expectation of conditioned RV	$\mathbf{a}$	$\bar{\mathbf{x}}$
Expectation of RV that we condition on	$\mathbf{b}$	$\mathbf{H}\bar{\mathbf{x}} = \bar{\mathbf{z}}$
Covariance of conditioned RV	$\mathbf{P}_{xx}$	$\bar{\mathbf{P}}$
Cross-covariance	$\mathbf{P}_{xy}$	$\bar{\mathbf{P}}\mathbf{H}^{\mathsf{T}}$
Covariance of RV that we condition on	$\mathbf{P}_{yy}$	$\mathbf{H}\bar{\mathbf{P}}\mathbf{H}^{\mathsf{T}}+\mathbf{R} = \mathbf{S}$

Based on this, we may furthermore identify $\boldsymbol{\mu}_{x|y}$ and $\mathbf{P}_{x|y}$ in Theorem 3.2.3 with the entities

$$\hat{\mathbf{x}} = \bar{\mathbf{x}} + \bar{\mathbf{P}}\mathbf{H}^{\mathsf{T}}(\mathbf{H}\bar{\mathbf{P}}\mathbf{H}^{\mathsf{T}}+\mathbf{R})^{-1}(\mathbf{z}-\mathbf{H}\bar{\mathbf{x}}) \tag{3.15}$$

$$\hat{\mathbf{P}} = \bar{\mathbf{P}} - \bar{\mathbf{P}}\mathbf{H}^{\mathsf{T}}(\mathbf{H}\bar{\mathbf{P}}\mathbf{H}^{\mathsf{T}}+\mathbf{R})^{-1}\mathbf{H}\bar{\mathbf{P}}. \tag{3.16}$$

We have thus specified all the entities involved in the quadratic forms corresponding to $p(\mathbf{z})$ and $p(\mathbf{x}|\mathbf{z})$, and it follows that

$$(\mathbf{x}-\bar{\mathbf{x}})^\mathsf{T}\bar{\mathbf{P}}^{-1}(\mathbf{x}-\bar{\mathbf{x}})+(\mathbf{z}-\mathbf{H}\mathbf{x})^\mathsf{T}\mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}\mathbf{x})=(\mathbf{z}-\bar{\mathbf{z}})^\mathsf{T}\mathbf{S}^{-1}(\mathbf{z}-\bar{\mathbf{z}})+(\mathbf{x}-\hat{\mathbf{x}})^\mathsf{T}\hat{\mathbf{P}}^{-1}(\mathbf{x}-\hat{\mathbf{x}})$$

In order to arrive at the final expressions in the theorem, we notice first that the term $\mathbf{W}=\bar{\mathbf{P}}\mathbf{H}^\mathsf{T}\mathbf{S}^{-1}$ is present in both (3.15) and (3.16). In the first of these, recognizing $\mathbf{W}$ leads to the expression $\hat{\mathbf{x}}=\bar{\mathbf{x}}+\mathbf{W}(\mathbf{z}-\mathbf{H}\bar{\mathbf{x}})$. In the second of these, recognizing $\mathbf{W}$ leads to $\hat{\mathbf{P}}=\bar{\mathbf{P}}-\mathbf{W}\mathbf{H}\bar{\mathbf{P}}=(\mathbf{I}-\mathbf{W}\mathbf{H})\bar{\mathbf{P}}$, and we have successfully separated (3.13) into the two quadratic forms on the right-hand-side of the product identity. ■

3.4 The canonical form

As an alternative to the moment-based parameterization (3.1), the multivariate Gaussian can also parameterized in the canonical form, which in the multivariate case becomes

$$\mathcal{N}(\mathbf{x};\mu,\mathbf{P})=\exp\left(a+\boldsymbol{\eta}^\mathsf{T}\mathbf{x}-\frac{1}{2}\mathbf{x}^\mathsf{T}\boldsymbol{\Lambda}\mathbf{x}\right) \tag{3.17}$$

where

$$\boldsymbol{\Lambda}=\mathbf{P}^{-1} \tag{3.18}$$
$$\boldsymbol{\eta}=\boldsymbol{\Lambda}\mu \tag{3.19}$$
$$a=-(1/2)n\ln(2\pi)-\ln|\boldsymbol{\Lambda}|+\boldsymbol{\eta}^\mathsf{T}\boldsymbol{\Lambda}\boldsymbol{\eta}. \tag{3.20}$$

The entity $\boldsymbol{\eta}$ is sometimes referred to as an information state. The entity $\boldsymbol{\Lambda}$ is known as the information matrix or precision matrix. The term "information" refers to the fact that $\boldsymbol{\Lambda}$, being the inverse of $\mathbf{P}$, is a measure of the opposite of uncertainty.

We saw in the previous section that in the moment-based parametrization, marginalization was easy while conditioning was more complicated. In canonical form, it is opposite.

Theorem 3.4.1 — Marginalization and conditioning in canonical form. If $\mathbf{x}$ and $\mathbf{y}$ have the joint distribution

$$p(\mathbf{x},\mathbf{y})\propto\exp\left(\begin{bmatrix}\boldsymbol{\eta}_a^\mathsf{T} & \boldsymbol{\eta}_b^\mathsf{T}\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}-\frac{1}{2}\begin{bmatrix}\mathbf{x}^\mathsf{T} & \mathbf{y}^\mathsf{T}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Lambda}_{xx} & \boldsymbol{\Lambda}_{xy}\\ \boldsymbol{\Lambda}_{xy}^\mathsf{T} & \boldsymbol{\Lambda}_{yy}\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\right) \tag{3.21}$$

then the marginal distribution of $\mathbf{y}$ is has potential vector $\boldsymbol{\eta}_*=\boldsymbol{\eta}_b-\boldsymbol{\Lambda}_{xy}^\mathsf{T}\boldsymbol{\Lambda}_{xx}^{-1}\boldsymbol{\eta}_a$ and information matrix $\boldsymbol{\Lambda}_*=\boldsymbol{\Lambda}_{yy}-\boldsymbol{\Lambda}_{xy}^\mathsf{T}\boldsymbol{\Lambda}_{xx}^{-1}\boldsymbol{\Lambda}_{xy}$, and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ has potential vector $\boldsymbol{\eta}_{x|y}=\boldsymbol{\eta}_a-\boldsymbol{\Lambda}_{xy}\boldsymbol{\eta}_b$ and information matrix $\boldsymbol{\Lambda}_{x|y}=\boldsymbol{\Lambda}_{xx}$.

Proof. First we show the marginalization, and then we use this to show the conditioning. For the marginalized density, the information matrix must be the inverse of the corresponding marginalized covariance matrix. This means that

$$\begin{aligned}\boldsymbol{\Lambda}_*=&\mathbf{P}_{yy}^{-1}\\ =&\left([\mathbf{0},\mathbf{I}]\begin{bmatrix}\mathbf{P}_{xx} & \mathbf{P}_{xy}\\ \mathbf{P}_{xy}^\mathsf{T} & \mathbf{P}_{yy}\end{bmatrix}\begin{bmatrix}\mathbf{0}\\ \mathbf{I}\end{bmatrix}\right)^{-1}=\left([\mathbf{0},\mathbf{I}]\begin{bmatrix}\boldsymbol{\Lambda}_{xx} & \boldsymbol{\Lambda}_{xy}\\ \boldsymbol{\Lambda}_{xy}^\mathsf{T} & \boldsymbol{\Lambda}_{yy}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{0}\\ \mathbf{I}\end{bmatrix}\right)^{-1}\\ =&\left([\mathbf{0},\mathbf{I}]\begin{bmatrix}\cdots & \cdots\\ \cdots & (\boldsymbol{\Lambda}_{yy}-\boldsymbol{\Lambda}_{xy}^\mathsf{T}\boldsymbol{\Lambda}_{xx}^{-1}\boldsymbol{\Lambda}_{xy})^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{0}\\ \mathbf{I}\end{bmatrix}\right)^{-1}\\ =&\boldsymbol{\Lambda}_{yy}-\boldsymbol{\Lambda}_{xy}^\mathsf{T}\boldsymbol{\Lambda}_{xx}^{-1}\boldsymbol{\Lambda}_{xy}.\end{aligned} \tag{3.22}$$

We have inverted the $\mathbf{\Lambda}$ matrix by means of the inverse block matrix formula from Appendix Q.1. Since the inverted matrix is to be pre- and post-multiplied by the matrix $[\mathbf{0},\mathbf{I}]$ we have simply written "…" instead of spelling out the irrelevant blocks. In order to express the marginalized potential, we first express the marginalized expectation in terms of the joint potential vector:

$$\mathbf{b} = [\mathbf{0},\mathbf{I}] \begin{bmatrix} \mathbf{\Lambda}_{xx} & \mathbf{\Lambda}_{xy} \\ \mathbf{\Lambda}_{xy}^\mathsf{T} & \mathbf{\Lambda}_{yy} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{\eta}_a \\ \boldsymbol{\eta}_b \end{bmatrix} \tag{3.23}$$

The marginalized potential vector can then be found as

$$\boldsymbol{\eta}_* = \mathbf{\Lambda}_* \mathbf{b} = \mathbf{\Lambda}_* (\mathbf{\Lambda}_*^{-1} \mathbf{\Lambda}_{yy} - \mathbf{\Lambda}_*^{-1} \mathbf{\Lambda}_{xy}^\mathsf{T} \mathbf{\Lambda}_{xx}^{-1} \boldsymbol{\eta}_a) = \boldsymbol{\eta}_b - \mathbf{\Lambda}_{xy}^\mathsf{T} \mathbf{\Lambda}_{xx}^{-1} \boldsymbol{\eta}_a. \tag{3.24}$$

The conditional Gaussian $p(\mathbf{x}|\mathbf{y})$ can then be found according to

$$\begin{aligned} p(\mathbf{x}|\mathbf{y}) =& \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{y})} \\ \propto& \exp\left(\boldsymbol{\eta}_a^\mathsf{T}\mathbf{x} + \boldsymbol{\eta}_b^\mathsf{T}\mathbf{y} - \frac{1}{2}\left(\mathbf{x}^\mathsf{T}\mathbf{\Lambda}_{xx}\mathbf{x} + \mathbf{x}^\mathsf{T}\mathbf{\Lambda}_{xy}\mathbf{y} + \mathbf{y}^\mathsf{T}\mathbf{\Lambda}_{xy}^\mathsf{T}\mathbf{x} + \mathbf{y}^\mathsf{T}\mathbf{\Lambda}_{yy}\mathbf{y} \right) \right. \\ & \left. - \left(\boldsymbol{\eta}_b - \mathbf{\Lambda}_{xy}^\mathsf{T}\mathbf{\Lambda}_{xx}^{-1}\boldsymbol{\eta}_a \right)^\mathsf{T}\mathbf{y} - \frac{1}{2}\mathbf{y}^\mathsf{T}\left(\mathbf{\Lambda}_{yy} - \mathbf{\Lambda}_{xy}^\mathsf{T}\mathbf{\Lambda}_{xx}^{-1}\mathbf{\Lambda}_{xy} \right)\mathbf{y} \right) \\ \propto& \exp\left(\boldsymbol{\eta}_a^\mathsf{T}\mathbf{x} - \mathbf{y}^\mathsf{T}\mathbf{\Lambda}_{xy}\mathbf{x} - \frac{1}{2}\mathbf{x}^\mathsf{T}\mathbf{\Lambda}_{xx}\mathbf{x} \right) \end{aligned} \tag{3.25}$$

The last proportionality is obtained by discarding all terms that do not contain $\mathbf{x}$, as they provide no information about the shape of the conditional distribution of $\mathbf{x}$. We see that $\boldsymbol{\eta}_a^\mathsf{T} - \mathbf{y}^\mathsf{T}\mathbf{\Lambda}_{xy}$ plays the role of the potential vector, while $\mathbf{\Lambda}_{xx}$ plays the role of the information matrix in the conditional distribution. ■

The canonical form is useful for several reasons. First, as shown in Theorem 3.4.1, conditioning is in general easier to do in canonical form than in the moment-based form. Since estimation often boils down to calcuating the marginal density of state, given the data, this is not unimportant. Furthermore, as already pointed out in (3.4), the information matrix equals the curvature of the logarithm of the Gaussian.

3.5 References and chapter remarks

The multivariate Gaussian is so important that extensive treatments can be found in many textbooks. From a statistical perspective, a standard reference is [41]. Readers more inclined towards machine learning, may prefer machine learning textbooks such as [53] or [6]. Sensor fusion textbooks such as [2] and [35] also contain all the standard stuff. The main purpose for the treatment in this book has been to present the most important results in manner as condensed and pedagogical as possible, while still provide rigorous proofs for all the results. It is therefore important to be aware that there is much more to read in the mentioned textbooks.

The strong focus on understanding the Gaussian as the exponential of a quadratic form is somewhat original to this book, although a similar perspective can also be found in [6]. The proof of Theorem 3.2.3 (marginalization and conditioning of Gaussians in moment-form) is due to Gary B. Huang [38], and goes along the lines of the proof in [53]. An alternative proof, formulated in terms of random variables, can be found in [41]. A third proof, which uses completion of the square, is found in [68]. The proof of the product identity follows the structure of the proof in [74]. Alternative proofs can be found in [65] and [50].

3.6 Exercises

Exercise 3.1 Let $\mathbf{A}$ and $\mathbf{B}$ be symmetric and invertible matrices of the same dimension. Show that

$$\mathbf{A} - \mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A} = (\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}.$$

Hint: The equality holds if and only if the inverse of the right-hand-side times the left-hand-side equals the identity matrix. ■

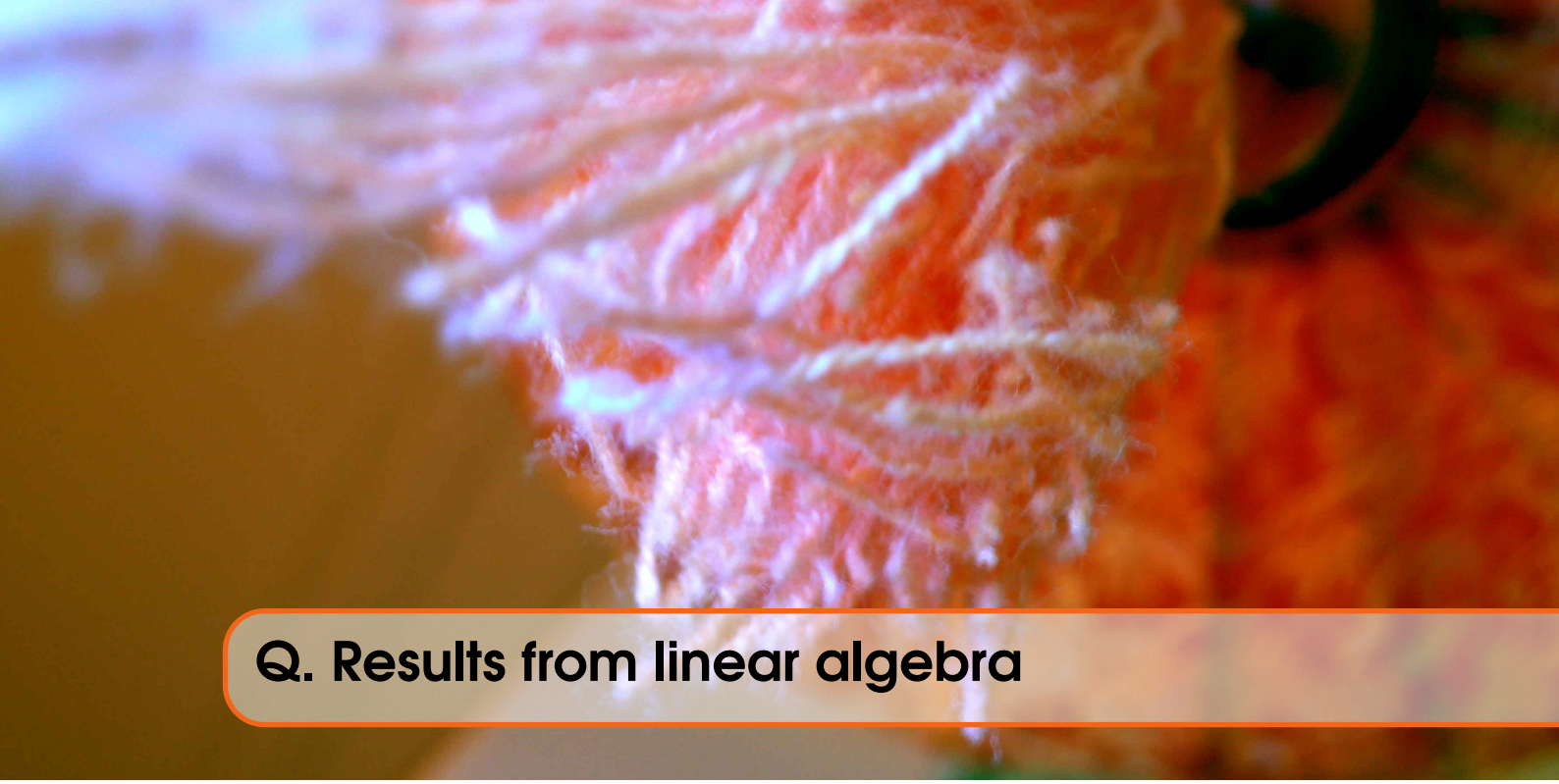
Exercise 3.2 Let $\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{w}$ where $\mathbf{x}$ and $\mathbf{w}$ are independent random vectors distributed according to $\mathcal{N}(\mathbf{x}; |\bar{\mathbf{x}}, \mathbf{P})$ and $\mathcal{N}(\mathbf{w}; |\mathbf{0}, \mathbf{R})$. Show that the joint distribution of $\mathbf{x}$ and $\mathbf{z}$ is

$$\mathcal{N}\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{z}\end{bmatrix}; \begin{bmatrix}\bar{\mathbf{x}}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{P} & \mathbf{P}\mathbf{H}^\mathsf{T}\\ \mathbf{H}\mathbf{P} & \mathbf{H}\mathbf{P}\mathbf{H}^\mathsf{T}+\mathbf{R}\end{bmatrix}\right).$$

■

Exercise 3.3 Derive the covariance formula $\hat{\mathbf{P}}^{-1} = \mathbf{H}^\mathsf{T}\mathbf{R}^{-1}\mathbf{H} + \bar{\mathbf{P}}^{-1}$. ■

Exercise 3.4 Let $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{P})$ be an n-dimensional RV. Show that $(\mathbf{x}-\mu)^\mathsf{T}\mathbf{P}^{-1}(\mathbf{x}-\mu)$ has a χ^2 distribution with n degrees of freedom.
Hint: You can build on relevant results from Examples 2.9, 2.10 and 3.4 as well as Theorem 3.2.2 and Exercise 2.2. ■



Q. Results from linear algebra

Q.1 The Schur complement and Boltz' inversion rule

Let the matrix $\mathbf{M}$ be given by

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}. \tag{Q.1}$$

The Schur complement of the block $\mathbf{D}$ in $\mathbf{M}$ is $\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}$, and the Schur complement of the block $\mathbf{A}$ in $\mathbf{M}$ is $\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}$. The Schur complement plays a central role when we invert matrices of the form (Q.1). With the purpose of obtaining a general formula for inverting such matrices, let us also introduce the matrices

$$\mathbf{L} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} \mathbf{I} & -\mathbf{B}\mathbf{D}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \quad \text{and} \quad \mathbf{N} = \begin{bmatrix} \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix}. \tag{Q.2}$$

Obviously, the goal here is to replace the inversion of a complicated matrix $\mathbf{M}$ with inversion of a simpler block-diagonal matrix $\mathbf{N}$ which involves the Schur complement.

Theorem Q.1.1 — Blockwise matrix inversion. With $\mathbf{M}$, $\mathbf{L}$, $\mathbf{U}$ and $\mathbf{N}$ defined in (Q.1) and (Q.2), we have the identity $\mathbf{M}^{-1} = \mathbf{L}\mathbf{N}^{-1}\mathbf{U}$.

Proof. By reverse engineering of the desired result, we find that $\mathbf{L}^{-1}\mathbf{M}^{-1} = \mathbf{N}^{-1}\mathbf{U}$ should hold, and furthermore that $\mathbf{M}\mathbf{L} = \mathbf{U}^{-1}\mathbf{N}$ should hold. The left-hand-side of this identity becomes

$$\mathbf{M}\mathbf{L} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{bmatrix} = \begin{bmatrix} \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} & \mathbf{B} \\ \mathbf{0} & \mathbf{D} \end{bmatrix}.$$

For the right-hand-side, we need to invert $\mathbf{U}$. We leave it as an exercise to the reader to show that

$$\mathbf{U}^{-1} = \begin{bmatrix} \mathbf{I} & \mathbf{B}\mathbf{D}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}.$$

Thus, the right-hand-side becomes

$$\mathbf{U}^{-1}\mathbf{N} = \begin{bmatrix} \mathbf{I} & \mathbf{B}\mathbf{D}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C} & \mathbf{B} \\ \mathbf{0} & \mathbf{D} \end{bmatrix}.$$

This concludes the proof. ■

Having established this important result, we can express the inverse of a block matrix in two ways which both are useful:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{bmatrix} \begin{bmatrix} (\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\mathbf{B}\mathbf{D}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \tag{Q.3}$$

$$= \begin{bmatrix} (\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & -(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & \mathbf{D}^{-1}+\mathbf{D}^{-1}\mathbf{C}(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \end{bmatrix} \tag{Q.4}$$

The first of these two equations (Q.3) is known as Aitken's block diagonalization formula, while the last equation (Q.4) is known as Boltz' rule for matrix inversion.

Q.2 The matrix inversion lemma

In Boltz' rule (Q.4) it can be noted that $\mathbf{A}$ and $\mathbf{D}$ are treated somewhat differently. This is because we chose to express $\mathbf{N}$ in terms of the Schur complement of $\mathbf{D}$. If we instead choose to work with the Schur complement of $\mathbf{A}$, symmetry dictates that we must arrive at

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1}+\mathbf{A}^{-1}\mathbf{B}(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \\ -(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & (\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \end{bmatrix}. \tag{Q.5}$$

The matrix inversion lemma, also known as the Woodbury identity, follows from equating the upper diagonal blocks in (Q.4) and (Q.5):

$$(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} = \mathbf{A}^{-1}+\mathbf{A}^{-1}\mathbf{B}(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}. \tag{Q.6}$$

The matrix inversion lemma is often invoked in derivations of the Kalman filter.

Q.3 Rodrigues' rotation formula

Rodrigues' formula was given in (12.1) and formed the basis for our development of rotation matrices and Euler angles. We repeat it here:

$$\boldsymbol{x}' = (1-\cos\alpha)(\boldsymbol{v}\cdot\boldsymbol{x})\boldsymbol{v} + \cos\alpha\boldsymbol{x} - \sin\alpha(\boldsymbol{x}\times\boldsymbol{n}). \tag{Q.7}$$

To derive this formula we start by decomposing the vector $\boldsymbol{x}$ into components parallel and perpendicular to the axis $\boldsymbol{v}$:

R. Matrix and vector derivatives

Bibliography

Books

[1] Y. Bar-Shalom and X. R. Li, *Multitarget-Multisensor Tracking: Principles and Techniques*. Storrs, CT, USA: YBS Publishing, 1995 (cited on pages 14, 63, 97, 99, 101, 109, 111, 124).

[2] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with Application to Tracking and Navigation*. Wiley, 2001 (cited on pages 13, 34, 45, 59, 60, 83, 88, 90, 99).

[4] Y. Bar-Shalom, P. K. Willett, and X. Tian, *Tracking and Data Fusion: A Handbook of Algorithms*. Storrs, CT, USA: YBS Publishing, 2011 (cited on pages 14, 124).

[5] T. D. Barfoot, *State estimation for robotics*. Cambridge University Press, 2018 (cited on page 14).

[6] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006 (cited on page 45).

[7] S. Blackman and R. Popoli, *Design and Analysis of Modern Tracking Systems*. Norwood, MA, USA: Artech House, 1999 (cited on pages 14, 142).

[17] S. Challa, M. R. Morelande, D. Musicki, and R. J. Evans, *Fundamentals of object tracking*. Cambridge University Press, 2011 (cited on pages 14, 109).

[23] O. Egeland and T. Gravdahl, *Modeling and Simulation for Automatic Control*. Trondheim, Norway: Marine Cybernetics, 2002 (cited on page 166).

[26] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*. Wiley, 1999 (cited on page 34).

[29] T. I. Fossen, *Handbook of Marine Craft Hydrodynamics and Motion Control*. Wiley, 2011 (cited on page 175).

[33] P. D. Groves, *Principles of GNSS, Inertial, and Multisensor Integrated Navigation Systems*, 2nd edition. Artech House, 2013 (cited on pages 14, 170).

[35] F. Gustafsson, *Statistical Sensor Fusion*. Lund, Sweden: Studentlitteratur, 2010 (cited on pages 13, 14, 45).

[41] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 5th edition. Prentice Hall, 2002 (cited on page 45).

[42] S. M. Kay, *Fundamentals of Statistical Signal Processing Volume I Estimation Theory*. Prentice-Hall, 1993 (cited on page 14).

[50] R. Mahler, *Statistical Multisource-Multitarget Information Fusion*. Norwood, MA, USA: Artech House, 2007 (cited on pages 14, 45, 102, 145, 148, 151, 157, 158).

[53] K. P. Murphy, *Machine Learning - A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012 (cited on page 45).

[60] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, 2002 (cited on pages 14, 27, 34, 55).

[63] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House, 2004 (cited on page 83).

[66] S. Särkkä, *Bayesian Filtering and Smooting*. Cambridge University Press, 2013 (cited on page 66).

[72] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*. MIT Press, 2006 (cited on page 14).

[73] D. Titterton and J. Weston, *Strapdown Inertial Navigation Technology*, 2nd edition, series Radar, Sonar and Navigation. IET, 2004 (cited on page 14).

[77] B. Vik, *Integrated Satellite and Inertial Navigation Systems*. Department of Engineering Cybernetics, NTNU, 2014 (cited on pages 14, 168).

[79] R. E. Walpole, R. H. Myers, and S. L. Myers, *Probability and Statistics for Engineers and Scientists*. Prentice Hall, 1998 (cited on page 34).

Articles

[3] Y. Bar-Shalom, S. S. Blackman, and R. J. Fitzgerald, "Dimensionless score function for multiple hypothesis tracking", *IEEE Transactions on Aerospace and Electronic Systems*, volume 43, number 1, pages 392–400, Jan. 2007. DOI: 10.1109/TAES.2007.357141 (cited on pages 141, 142).

[8] J.-L. Blanco, "A tutorial on se(3) transformation parameterizations and on-manifold optimization", MAPIR: Grupo de Percepción y Robótica, Universidad de Málaga, Tech. Rep., Oct. 2014 (cited on page 168).

[9] H. A. P. Blom and E. A. Bloem, "Probabilistic Data Association Avoiding Track Coalescence", *IEEE Transactions on Automatic Control*, volume 45, number 2, pages 247–259, Feb. 2000, ISSN: 0018-9286. DOI: 10.1109/9.839947 (cited on page 122).

[10] E. Brekke and M. Chitre, "The multiple hypothesis tracker derived from finite set statistics", in *20th International Conference on Information Fusion*, Xi'An, China, Jul. 2017, pages 1–8 (cited on pages 128, 155).

[11] E. Brekke, "Clutter Mitigation for Target Tracking", PhD thesis, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, Jun. 2010 (cited on page 63).

[12] E. Brekke and M. Chitre, "A multi-hypothesis solution to data association for the two-frame SLAM problem", *The International Journal of Robotics Research*, volume 34, number 1, pages 43–63, Jan. 2015. DOI: 10.1177/0278364914545674 (cited on page 183).

[13] E. Brekke, O. Hallingstad, and J. Glattetre, "The signal-to-noise ratio of human divers", in *Proceedings of OCEANS'10*, Sydney, Australia, May 2010 (cited on pages 64, 159).

[14] ——, “Tracking small targets in heavy-tailed clutter using amplitude information”, *IEEE Journal of Oceanic Engineering*, volume 35, number 2, pages 314–329, May 2010 (cited on pages 101, 123).

[15] ——, “The Modified Riccati Equation for Amplitude-Aided Target Tracking in Heavy-Tailed Clutter”, *IEEE Transactions on Aerospace and Electronic Systems*, volume 47, number 4, pages 2874–2886, Oct. 2011. DOI: 10.1109/TAES.2011.6034670 (cited on page 123).

[16] ——, “Improved target tracking in the presence of wakes”, *IEEE Transactions on Aerospace and Electronic Systems*, volume 48, number 2, pages 1005–1017, Apr. 2012 (cited on page 123).

[18] D. Clark, B. Ristic, and B.-N. Vo, “PHD filtering with target amplitude feature”, in *Information Fusion, 2008 11th International Conference on*, Cologne, Jun. 2008, pages 1–7 (cited on page 101).

[19] J. L. Crassidis and F. L. Markley, “Attitude estimation using modified rodrigues parameters”, in *Proceedings of the Flight Mechanics/Estimation Theory Symposium*, Greenbelt, MD, USA, 1996, pages 71–83 (cited on page 161).

[20] R. Danchick and G. E. Newnam, “Reformulating Reid’s MHT method with generalised Murty K-best ranked linear assignment algorithm”, *IEE Proceedings - Radar, Sonar and Navigation*, volume 153, pages 13–22, Feb. 2006 (cited on page 117).

[21] S. Davey, M. G. Rutten, and B. Cheung, “A Comparison of Detection Performance for Several Track-before-Detect Algorithms”, *EURASIP Journal on Advances in Signal Processing*, volume 2008, 2008. DOI: 10.1155/2008/428036 (cited on page 142).

[22] S. Deb, M. Yeddanapudi, K. Pattipati, and Y. Bar-Shalom, “A generalized S-D assignment algorithm for multisensor-multitarget state estimation”, *IEEE Transactions on Aerospace and Electronic Systems*, volume 33, number 2, pages 523–538, Apr. 1997, ISSN: 0018-9251. DOI: 10.1109/7.575891 (cited on pages 138–140).

[25] A. L. Flåten and E. Brekke, “Rao-blackwellized particle filter for turn rate estimation”, in *Proceedings of IEEE Aerospace Conference*, Big Sky, MT, USA, Mar. 2017 (cited on pages 83, 101).

[27] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, “On-manifold preintegration for real-time visual–inertial odometry”, *IEEE Transactions on Robotics*, volume 33, number 1, pages 1–21, Feb. 2017, ISSN: 1552-3098. DOI: 10.1109/TRO.2016.2597321 (cited on page 161).

[28] T. E. Fortmann, Y. Bar-Shalom, and M. Scheffe, “Sonar tracking of multiple targets using joint probabilistic data association”, *IEEE Journal of Ocean Engineering*, volume 3, pages 173–184, 1983 (cited on pages 111, 124).

[30] B. Gade, M. Kloster, and M. Aronsen, “Non-elliptical validation gate for maritime target tracking”, in *2018 21st International Conference on Information Fusion (FUSION)*, Jul. 2018, pages 1301–1308. DOI: 10.23919/ICIF.2018.8455282 (cited on page 109).

[31] K. Gade, “Integrering av treghetsnavigasjon i en autonom undervannsfarkost”, FFI, Tech. Rep. 97/03179, 1997 (cited on pages 170, 175).

[32] ——, “Inertial navigation - theory and applications”, PhD thesis, NTNU, 2018 (cited on page 175).

[34] M. Guignard, “Lagrangean relaxation”, *Top*, volume 11, number 2, 2003 (cited on page 138).

[36] G. Hendeby, “Performance and implementation aspects of nonlinear filtering”, PhD thesis, Linköping University, 2008 (cited on page 83).

[37] P. Horridge and S. Maskell, "Real-time tracking of hundreds of targets with efficient exact JPDAF implementation", in *9th International Conference on Information Fusion*, Florence, Jul. 2006 (cited on page 117).

[39] G. P. Huang, A. I. Mourikis, and S. I. Roumeliotis, "Observability-based rules for designing consistent EKF SLAM estimators", *The international Journal of Robotics Research*, volume 29, number 5, pages 502–528, Apr. 2010, ISSN: 02783649. DOI: 10.1177/0278364909353640 (cited on pages 185, 186).

[43] A. Kong, J. S. Liu, and W. H. Wong, "Sequential imputations and bayesian missing data problems", *Journal of the American Statistical Association*, volume 89, number 425, 1994 (cited on page 77).

[44] D. Lerro and Y. Bar-Shalom, "Interacting Multiple Model Tracking with Target Amplitude Feature", *IEEE Transactions on Aerospace and Electronic Systems*, volume 29, number 2, pages 494–509, Apr. 1993 (cited on page 101).

[45] ——, "Tracking with debiased consistent converted measurements versus EKF", *IEEE Transactions on Aerospace and Electronic Systems*, volume 29, number 3, pages 1015–1022, Jul. 1993, ISSN: 0018-9251 (cited on page 63).

[46] X. R. Li, "Tracking in clutter with strongest neighbor measurements. i. theoretical analysis", *IEEE Transactions on Automatic Control*, volume 43, number 11, pages 1560–1578, Nov. 1998 (cited on page 95).

[48] E. Liland, "AIS aided multi hypothesis tracker", Master's thesis, NTNU, Jun. 2017 (cited on page 137).

[49] M. Longbin, S. Xiaoquan, Z. Yiyu, S. Z. Kang, and Y. Bar-Shalom, "Unbiased Converted Measurements for Tracking", *IEEE Transactions on Aerospace and Electronic Systems*, volume 34, number 3, pages 1023–1027, 1998 (cited on page 63).

[51] M. Montemerlo, "FastSLAM: A factored solution to the simultaneous localization and mapping problem with unknown data association", PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, Jun. 2003 (cited on page 187).

[52] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit, "FastSLAM: A factored solution to the simultaneous localization and mapping problem", in *Proceedings of AAAI-02*, Edmonton, AL, Canada, 2002, pages 593–598 (cited on page 187).

[54] D. Musicki, R. Evans, and S. Stankovic, "Integrated Probabilistic Data Association", *IEEE Transactions on Automatic Control*, volume 39, number 6, pages 1237–1241, 1994 (cited on pages 102, 103, 107, 109).

[55] D. Musicki and S. Suvorova, "Tracking in clutter using IMM-IPDA-based algorithms", volume 44, number 1, pages 111–126, Jan. 2008, ISSN: 00189251 (cited on page 124).

[56] J. Neira and J. D. Tardós, "Data association in stochastic mapping using the joint compatibility test", *IEEE Transactions on Robotics and Automation*, volume 17, number 6, pages 890–897, Dec. 2001, ISSN: 1042296X. DOI: 10.1109/70.976019 (cited on page 183).

[57] P. C. Niedfeldt, K. Ingersoll, and R. W. Beard, "Comparison and analysis of Recursive-RANSAC for multiple target tracking", *IEEE Transactions on Aerospace and Electronic Systems*, volume 53, number 1, pages 461–476, Feb. 2017, ISSN: 0018-9251 (cited on pages 123, 124, 141).

[58] P. C. Niedfeldt, "Recursive-ransac: A novel algorithm for tracking multiple targets in clutte", PhD thesis, Brigham Young University, 2014 (cited on page 141).

[59] J. Olofsson, E. Brekke, T. I. Fossen, and T. A. Johansen, "Spatially indexed clustering for scalable tracking of remotely sensed drift ice", in *Proceedings of IEEE Aerospace Conference*, Big Sky, MT, USA, Mar. 2017 (cited on page 117).

[62] D. Reid, "An algorithm for tracking multiple targets", *IEEE Transactions on Automatic Control*, volume 24, number 6, pages 843–854, Dec. 1979 (cited on pages 125, 128).

[64] S. Roumeliotis, G. Sukhatme, and G. A. Bekey, "Circumventing dynamic modeling: Evaluation of the error-state Kalman filter applied to mobile robot localization", in *Proceedings of ICRA*, volume 2, 1999, 1656–1663 vol.2 (cited on page 175).

[65] D. J. Salmond, "Tracking in uncertain environments", Royal Aerospace Establishment, UK, Tech. Rep. AW 121, Sep. 1989 (cited on page 45).

[67] T. Schön, "On computational methods for nonlinear estimation", PhD thesis, Linköping University, 2003 (cited on page 83).

[69] R. Smith, M. Self, and P. Cheeseman, "Estimating uncertain spatial relationships in robotics", *Autonomous robot vehicles*, pages 167–193, 1990 (cited on page 178).

[71] E. F. Wilthil, A. L. Flaten, and E. F. Brekke, "A target tracking system for asv collision avoidance based on the pdaf", English, in *Sensing and Control for Autonomous Vehicles*, P. Fossen and Nijmeijer, Eds., volume 474, Alesund, Norway: Springer, 2017, pages 269–288 (cited on pages 61, 64).

[74] L.-C. N. Tokle, "Multi target tracking using random finite sets with a hybrid state space and approximations", Master's thesis, NTNU, Sep. 2018 (cited on page 45).

[75] N. Trawny and S. I. Roumeliotis, "Indirect kalman filter for 3d attitude estimation - a tutorial for quaternion algebra", MARS-LAB, Tech. Rep., 2005 (cited on page 175).

[76] C. Van Loan, "Computing integrals involving the matrix exponential", *IEEE Transactions on Automatic Control*, volume 23, number 3, pages 395–404, 1978 (cited on page 59).

[78] B.-N. Vo, S. Singh, and A. Doucet, "Sequential Monte Carlo methods for multitarget filtering with random finite sets", *IEEE Transactions on Aerospace and Electronic Systems*, volume 41, number 4, pages 1224–1245, Oct. 2005, ISSN: 0018-9251. DOI: 10.1109/TAES.2005.1561884 (cited on page 151).

[80] P. Willett, Y. Ruan, and R. Streit, "PMHT: Problems and Some Solutions", *IEEE Transactions on Aerospace and Electronic Systems*, volume 38, number 3, pages 738–754, Jul. 2002, ISSN: 0018-9251. DOI: 10.1109/TAES.2002.1039396 (cited on page 142).

[81] J. Williams, "An efficient, variational approximation of the best fitting multi-Bernoulli filter", *IEEE Transactions on Signal Processing*, volume 63, number 1, pages 258–273, Jan. 2015, ISSN: 1053-587X. DOI: 10.1109/TSP.2014.2370946 (cited on page 122).

[82] ——, "Marginal multi-Bernoulli filters: RFS derivation of MHT, JIPDA, and association-based MeMBer", *IEEE Transactions on Aerospace and Electronic Systems*, volume 51, number 3, pages 1664–1687, Jul. 2015 (cited on pages 109, 154).

[83] J. Williams and R. Lau, "Approximate evaluation of marginal association probabilities with belief propagation", *IEEE Transactions on Aerospace and Electronic Systems*, volume 50, number 4, pages 2942–2959, Oct. 2014 (cited on page 117).

[84] E. F. Wilthil, E. Brekke, and O. B. Asplin, "Track initiation for maritime radar tracking with and without prior information", in *Proc. Fusion*, Cambridge, UK, Jul. 2018 (cited on page 106).

[85] Y. Xia, K. Granström, L. Svensson, and Á. F. García-Fernández, "An implementation of the Poisson Multi-Bernoulli mixture trajectory filter via dual decomposition", in *2018 21st International Conference on Information Fusion (FUSION)*, Jul. 2018, pages 1–8. DOI: 10.23919/ICIF.2018.8455236 (cited on page 138).

Sources and resources from the web

[24] M. A. T. Figueiredo, "Lecture notes on Bayesian estimation and classification", Instituto de Telecomunicacoes, Portugal., Oct. 2004 (cited on page 34).

[38] G. B. Huang, "Conditional and marginal distributions of a multivariate Gaussian", Accessed 17th of June 2017, Feb. 2010, [Online]. Available: https://gbhqed.wordpress.com/2010/02/21/conditional-and-marginal-distributions-of-a-multivariate-gaussian (cited on pages 45, 188).

[40] T. A. Johansen and T. I. Fossen, "The eXogenous Kalman filter (XKF)", Submitted to International Journal of Control, 2015 (cited on page 186).

[47] E. Liland, "An ILP approach to multi hypothesis tracking", Specialization project at NTNU, Dec. 2016 (cited on page 138).

[61] J. Pedersen, "Surveillance of the channel", Specialization Project at NTNU, Dec. 2017 (cited on page 122).

[68] T. B. Schön and F. Lindsten, "Manipulating the Multivariate Gaussian Density", Accessed 17th of June 2017, Jan. 2011, [Online]. Available: user.it.uu.se/~thosc112/pubpdf/schonl2011.pdf (cited on page 45).

[70] J. Solà, "Quaternion kinematics for the error-state KF", Mar. 2015, [Online]. Available: http://www.iri.upc.edu/people/jsola/JoanSola/objectes/notes/kinematics.pdf (cited on pages 14, 163, 166, 170–172, 175).