Probabilistic inference, Gaussians, quadratics, and the Kalman filter

Nathan Ratliff

Abstract

Gaussian distributions are perhaps the most important distribution you'll ever encounter, not because they represent everything well (they're usually, by themselves, poor approximations to complex systems), but because they're one of the only high-dimensional distributions we can handle analytically. And because of that, many approaches to more complicated problems revolve around reducing the problems to Gaussian approximations or sequences of Gaussian approximations to leverage the tractable algebra of Gaussian inference, which, to a large extent, itself reduces to linear algebra. That's not to say Gaussian manipulations easy— in many ways the algebra can be tedious— but having a thorough understanding of their properties, their relation to quadratic functions and linear systems, and their manipulation in the context of probabilistic inference for such queries as the transition and observation updates of a Kalman filter is crucial for a strong foundation for further study within the uncertainty-laden domain of state-estimation, localization, mapping, and sensor processing in mobile robotics.

1 Introduction

The calculus of probability can be tricky, even for the mathematically and logically minded engineers who frequent technical fields such as robotics. This document is designed to build up some of the fundamental ideas central to reasoning about uncertainty from an intuitive and geometric standpoint in order to develop a strong fundamental understanding of the Gaussian distributions and their manipulations. It culminates in a discussion of the Kalman filter and some notes on the relationship between inference with Gaussian random variables and ideas in optimization.

This material caters to budding researchers in robotics, which means that there's a baseline of technical fluency required to understand these ideas. For instance, I discuss multi-dimensional spaces, vectors, matrices and other such foundational concepts without much fanfare and assume these concepts are familiar. That said, I've found that geometric perspectives on potentially familiar objects (e.g. positive definite matrices and their associated Eigenspectrum) are often emphasized less in introductory courses than they should be. I spend

quite a bit of time developing these geometric perspectives here since they offer insight into the interaction between a Gaussian's covariance and it's geometry. I assume little knowledge of probability theory and try to develop much of what I use from basic principles or at least strong intuition. As such, some of this material may go too slow at times for those already familiar with the ideas. Certainly, skip sections that are boring and come back if things don't make sense later on.

The appendices in this document aren't necessarily auxiliary sections in the traditional sense (i.e. they're not just intended to be reference sections for background material), but they're more in depth offshoots into areas important to the discussion but who's inclusion in the document's core would have unduly impeded the flow. The first appendix discusses the geometry of positive definite matrices and the associated geometry of the covariance matrix of a single multivariate random variable, the second generalizes those ideas to covariances between two different multivariate random variables, and the third briefly reviews a basic formula central to the manipulation of Gaussians and the derivation of the Kalman filter and expands on its geometric interpretation in the context of the results presented in the previous two appendices.

Linear systems are some of the only systems we can directly manipulate algebraically and understand in depth. When analyzing non-linear systems, we often invoke linearizations (Taylor expansions or tangent spaces of differentiable manifolds, for instance) in order to reduce them locally back to the better understood space of linear systems. In that sense, one of the best understood functions in optimization is the quadratic function, since setting it's gradient to zero to solve for the critical point (minimum, maximum, or saddle-point) results in solving well understood linear systems. Quadratic functions, in the context of optimization, are therefore essentially just linear systems.

The Gaussian distribution is built around the quadratic function (it's inversely proportional to an exponentiated quadratic), and, because of that, it's largely governed by principles of linear algebra. Manipulating Gaussians, as we'll see, is an exercise in manipulating and analyzing linear systems, making Gaussian distributions one of the only closed-form continuous distributions we can directly manipulate in practice. Many non-Gaussian systems are solved or approximated by reducing them to a single Gaussian, a series of Gaussians en-route toward an iterative solution, or collection (mixture) of Gaussians. The Gaussian distribution is the hammer of probabilistic inference, system modeling, and analysis; a thorough understanding of its geometry and role in inference is critical.

We start from the basics, introducing first random variables and core statistical quantities such as expected values and covariances. We then move into a discussion of the geometry of Gaussian distributions (and more generally in the appendix, how the Eigenvectors and Eigenvalues of positive definite matrices expose their geometry through the Eigen-decomposition of the matrix). The discussion culminates in a simple derivation of the Kalman filter and some observations on the relationship between optimization and probabilistic inference.

2 Random variables: Expectations and Covariances

This section starts with a very broad and basic discussion of what a random variable is. Readers already familiar with the concept should skip directly to Sections 2.2 and 2.3 where we define the expected value (mean) and covariance of a random variable and examine how these quantities transform under affine transforms of the underlying random variables.

2.1 Random variables

Random variables are a probabilistic extension of the basic idea of variables in mathematics. If you've ever tried teaching a young student the idea of variables when they first encounter them in algebra, or if you can remember back to your own first encounters with them, they're not necessarily the easiest concept to grasp initially. Once you lock onto the right perspective, then they're trivial and you wonder why you ever had difficulty (or why your students ever had difficulty), but communicating that concept is at times tricky. The same thing holds with random variables in probability theory. Once you know what a random variable is, it the idea of random variables is natural and there's seemingly no reason why anyone should ever be confused by them. But until you have a solid picture in your mind about what these things are, understanding probabilistic or statistical principles can seem daunting. So this section is intended to ensure that everyone's on the same page with regard to what a random variable is.

Regular variables in mathematics, such as x in the simple algebraic expression $5x + 7$, are symbols that means "any number". In a sense, they represent all numbers simultaneously; you haven't chosen any particular number for it to be, so the entire real number line is fair game. In the same way, a *random* variable, often denoted by a capitol letter X, also represents any number in a domain (e.g. all of the real numbers, or some subset thereof), but with a twist. Not all numbers are equally likely!

You can think of X as a giant bucket of items. If we reach in and pull something out, some of the items will be more likely to come out than others. For instance, all of the real numbers might be in the bucket[1] in which case, when we reach in and grab something, the something we pull out is a real number. We call this process *sampling* from the random variable.

The most interesting case is when not all items in the bucket are equally likely. For instance, suppose we repeatedly pull numbers from the bucket of real values and keep track of where they fall on the real number line. Perhaps most numbers are centered around the value 5.0 and maybe 90% of them are between

[1] Technically, then, we'd have to think of it as an infinitely long string coiled up in the bucket, and pulling something from the bucket would mean pulling some infinitesimally small snippet from the string, but that's perhaps where the analogy breaks down. Mathematical abstractions were invented to get around physicality problems like this in thought experiments, so for our purposes, we can just say that the bucket has stuff in it and we can reach in and pull individual elements out.

3.0 and 7.0. That gives us a lot of information about the contents of the bucket and allows us to predict with high likelihood what's we might pull out of the bucket next if we reach in again. This information tells us something about the *distribution* of values we see coming from the bucket.

Formally, we model this distribution of numbers as a real-valued function $p(x)$ that maps some domain $\mathcal{X}$ (in this case, the real number line) to the positive (or non-negative) real numbers $p : \mathcal{X} \to \mathbb{R}_+$. We won't go into depth about the rigorous requirements these functions, beyond to say that in order for them to be meaningful and consistent, they need to integrate to 1 ($\int_{\mathcal{X}} p(x)dx = 1$), and the probability that a number x sampled from X falls into a set $\mathcal{S}$ is the integral of the density over just that set $\int_{\mathcal{S}} p(x)dx$.

In analogy to physics (ideas like mass density and total mass), we call $p(x)$ the *probability density function* since it tells us the amount of *probability mass* per unit volume that sits at a point. The integral of this probability density then tells us how much total *probability mass* falls within that region. High density over a region results in high probability within that region (i.e. high likelihood of pulling out a number that falls into that region in our example), and, vice versa, low density within a region leads to very low likelihood of ever seeing items that fall within that region when sampling. Since the density integrates to 1 we can say that the total probability mass over the entire domain is 1, and that every integral over a subregion we perform gives us the proportion of the probability mass that falls into that subregion. In a precise sense, this means that calculating $\int_{\mathcal{S}} p(x)dx$ for $\mathcal{S} \subset \mathcal{X}$ tells us the proportion of the time random samples from X will fall specifically into the subregion $\mathcal{S}$.

Often we're interested in how multiple random quantities relate to one another. In those cases, there may be multiple random variables X_i that we stack into a random vector

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}.$$

We refer to X as a *multivariate* random variable. In most of what we see below we'll assume that we're dealing with multivariate random variables.

A typical operation of interest within probability is the transformation of these random variables by a function, such as $Y = f(X)$, where f is some function $f : \mathcal{X} \to \mathcal{Y}$. It's sometimes easiest to think of transformations like this in terms of the process of sampling. If we sample a value from X (e.g. pull an item out of the bucket), and then transform by f into the new space $\mathcal{Y}$, then we effectively get a sample from $\mathcal{Y}$. If we do that repeatedly, we get a distribution of samples within $\mathcal{Y}$ that we can represent using a probability density function over $\mathcal{Y}$. The distribution over Y is related to the distribution over X in that it is the distribution over X warped by the mapping f, but in general the two can be drastically different in terms of their shape and properties.

This document is most interested in affine transformations of the form

$$Y = f(X) = AX + b$$

for some matrix A and vector b. As we'll see below, if X is a Gaussian random variable (i.e. has a Gaussian density, which we define precisely in Section 3), then Y, created as an affine transform of X, is also a Gaussian random variable. This is one of the properties that makes Gaussians such a pleasure to work with.

2.2 Expected values and affine transformations

Suppose there's some random happening that we might be able to model with a random variable X. For instance, the random variable might describe features of sheep that live on a particular ranch. And suppose we can take a set of measurements of each of these sheep that tell us, say, how much each one ultimately weighs when grown up and how much wool we get from it. Then one question we might want to answer is how much wool, on average, would we expected to get from each sheep at the end of the year.

To make this precise, we can say that our multivariate random variable X has a probability density $p(x)$ modeling the distribution of different types of sheep we expect to find in a given year. We can represent the quantity in question, the amount of wool we yield from each sheep, as a measurement function $f : \mathcal{X} \to \mathbb{R}^k$. In this case k is just 1, but in more general settings you can imagine there being multiple quantities we want to measure.

Now we have the tools to formalize this question of expected measurement. We say that the expected value of $f(X)$ is given by

$$E[f(X)] = \int f(x)\ p(x)dx. \tag{1}$$

Since $\int p(x)dx = 1$, we can think of this as the weighted average over all x of the values $f(x)$ using weights $p(x)$. The more $p(x)$ is peaked around a particular value, the more this weighted average tends to represent that value.[2] Typically, when we say "*the* expected value", we're referring to the expected value of the domain itself: if we sample repeatedly from $p(x)$, then what value do we expect to see on average. In this case, our measurement is the identity $f(x) = x$, and we have

$$E[X] = \int x\ p(x)dx.$$

The expected value of a random variable itself is often called the "mean".

[2] In the limiting case, where the distribution becomes infinitely peaked around a particular value x_0, this actually becomes the value $f(x_0)$. That strange object who is zero everywhere except at x_0 where it's infinitely peaked (and always integrates to 1), is known formally as a Dirac delta function and denoted $\delta_{x_0}(x)$. It crops up in a number of contexts throughout statistics (especially nonparametric statistics), physics, and functional analysis, but we won't discuss it further in document. We generally assume that the density function is everywhere finite.

For our purposes, what's interesting about the expected value defined in Equation 1 is how it changes under affine transformations. If a random variable Y is defined as affine transformation of another random variable $Y = AX + b$, where A is a matrix and b is a vector (both non-random), then the expected value of this new random variable is simply the same affine transformation of the original expected value:

$$\begin{aligned} E[Y] = E[AX + b] &= \int (Ax + b)p(x)dx \\ &= A\left(\int x\ dx\right) + b \\ &= A\ E[X] + b. \end{aligned}$$

2.3 Covariance matrices and affine transformations

Another thing we might ask if we have two multivariate random variables is how the dimensions of one vary with the dimensions of the other.[3] For instance, if X is a random variable over an m-dimensional space and Y is a random variable over an n-dimensional space, we often wonder whether the ith dimension X is high when the jth dimension of Y is large. Or in the opposite sense, whether the ith dimension of X is low when the jth dimension of Y is large. Or maybe knowing the ith dimension of X tells us nothing about the jth dimension of Y, in which case that notion of independence between the two variables is often useful to know as well.

More concretely, suppose that the dimensions the random variable X list out meteorological (weather) data measurements of a particular region, and the random variable Y's dimensions tell us the percentage of time it snows, rains, is sunny, etc. there. If ith dimension of X represents the humidity, and the jth dimension of Y is the percentage of time throughout the day it's thunderstorming, we'd expect the ith dimension of X to be large (high humidity) whenever the jth dimension of Y is large (high percentage of thunderstorms). On the other hand, we expect the opposite to be true between Y_j and another dimension X_k if that kth dimension represents the amount of sunshine that location's getting. Each dimension of X is potentially related to each dimension of Y—the covariance matrix formalizes these intuitive interconnections.

2.3.1 Covariance formalities and intuition

The mathematical definition of *covariance* is

$$\text{Cov}(X, Y) = E\left[(X - E[X])(Y - E[Y])^T\right]. \tag{2}$$

To the uninitiated, this expression can look daunting, but once you've stared at it long enough, it's actually pretty intuitive. We'll go through it piece by piece.

[3] The two random variables might actually be the same random variable, in which case we want to know how each of the dimensions vary with one-another.

First subtracting off the mean essentially just places us within a coordinate system centered around the mean value. Rather than considering values of X relative to some arbitrary zero point, $\widetilde{X} = X - E[X]$ considers how the value deviates from the expected value. Said colloquially, we can say we consider how the value deviates from what we'd expect.[4] In these coordinates, the covariance is simply looking the expected value of the outer products: $E[\widetilde{X}\widetilde{Y}^T]$.

Lets take a look more closely at what the expected value of the outer product is

$$E[\widetilde{X}\widetilde{Y}^T] = \begin{pmatrix} E[\widetilde{X}_1\widetilde{Y}_1] & E[\widetilde{X}_1\widetilde{Y}_2] & \cdots & E[\widetilde{X}_1\widetilde{Y}_n] \\ E[\widetilde{X}_2\widetilde{Y}_1] & E[\widetilde{X}_2\widetilde{Y}_2] & \cdots & E[\widetilde{X}_2\widetilde{Y}_n] \\ \vdots & \vdots & \ddots & \vdots \\ E[\widetilde{X}_m\widetilde{Y}_1] & E[\widetilde{X}_m\widetilde{Y}_2] & \cdots & E[\widetilde{X}_n\widetilde{Y}_n] \end{pmatrix}. \tag{3}$$

Each of these entries is of the form $E[\widetilde{X}_i\widetilde{Y}_j]$. If the values (deviations from the expected value) $\widetilde{X}_i$ and $\widetilde{Y}_j$ tend to both be largely positive or largely negative when the other is, then this expected value is going to be large. Conversely, if the opposite is true and one is typically largely positive when the other is largely negative and vice versa, then this expected value is going to be largely negative. On the other hand, if there's no rhyme or reason to the relationship between values of $\widetilde{X}_i$ and $\widetilde{Y}_j$, i.e. sometimes they're the largely positive or negative together and sometimes they're the opposite, then the positive products will cancel the negative products in the expectation and the overall expected value will end up close to zero (if not exactly zero).

The **Covariance Matrix**, which is given explicitly component-wise in Equation 3 and defined in Equation 2 is literally a matrix whose entries specify the extent to which the dimensions of random variables vary with one another. We often talk about the covariance matrix of a single random variable. When we do, we formally mean, the covariance between the random variable and itself: $\text{Cov}(X) = \text{Cov}(X, X)$. From the definition above, we see that

$$\text{Cov}(X) = E[\widetilde{X}\widetilde{X}^T]$$

is a symmetric matrix whose ijth entry tells us the degree to which dimension i varies with dimension j.

2.3.2 The affect of linear transformations

Now that we have these concrete definitions, we can ask what the Covariance Matrix of a linearly transformed random variable $Y = AX + b$ is. Here, again, as in Section 2.2, both A and b are non-random, i.e. A is just a matrix and b is

[4] This statement is only partially correct in that when the distribution is multimodal, the expected value may be a point in a desert land of low-probability between local maxima, making the mean potentially far from "what we'd expect" to see from the distribution. That said, the phrase is a nice one and intuitive enough warrant repeating.

just a vector. Writing out the definition directly gives

$$\begin{aligned}\text{Cov}(Y) &= E[(Y - E[Y])(Y - E[Y])^T] \\ &= E[(A(X - E[X]))(A(X - E[X]))^T] \\ &= A\ E[\widetilde{X}\widetilde{X}^T]A^T = A\text{Cov}(X)A^T.\end{aligned}$$

In other words, the covariance matrix of X, once pushed through the affine transformation $Y = AX + b$ becomes sandwiched between A and A^T. This transformation is often called a *similarity transform* of the covariance matrix by A.

To summarize, for any random variable, linear transformations $X \to AX + b$ behave very nicely in terms of both expectation and covariance computations.

1. The **expected value** is just pushed through the same linear transformation that's transforming the random variable $E[X] \to A\ E[X] + b$, and
2. the **covariance matrix** becomes similarity transformed $\text{Cov}(X) \to A\ \text{Cov}(X)A^T$.

3 Gaussian distributions and quadratic functions

Gaussian distributions are probability distributions entirely specified by their mean and covariance. In fact, there's a rigorous sense in which they're actually the most uncertain distributions (i.e. they make the least assumptions) that have those particular "statistics."[5]

The general form of multi-dimensional Gaussian, denoted $\mathcal{N}(\mu, \Sigma)$, where μ is the mean and Σ is the covariance matrix, is given by

$$p(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left\{-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right\},$$

where $|\Sigma|$ denotes the determinant of the covariance and n is the dimensionality of the random variable. Already, we can see a close connection between quadratic functions and the Gaussian distribution. The Gaussian distribution is nothing more than a distribution inversely proportional to an exponentiated quadratic. When the quadratic function gets really big, the probability gets exceedingly small. And when the quadratic achieves is minimum value, in this

[5]The term *statistic* is commonly used to refer to values computed to express properties of the distribution. In this case, we often say that the mean and covariance, together, form a *sufficient statistic* for the Gaussian distribution. Additionally, the notion of maximal uncertainty in this discussion comes from a principle known as the principle of maximum entropy. We can write down a mathematical optimization problem that explicitly encodes the search for a probability distribution (out of the space of all probability distributions) that 1. has precisely the mean and covariance we're considering, and 2. has the maximal *entropy*. (Entropy here is just an information theoretic quantity measuring uncertainty. Formally, it's $\mathcal{H}(X) = -\int p(x) \log p(x) dx$.) If we then solve that mathematical optimization problem analytically, we can show that the unique solution is the Gaussian distribution.

case specifically at the mean $x^* = \mu$, the probability distribution achieves its maximal value. This is actually a very useful property of the Gaussian: the mean is the maximizer of the distribution and the minimizer of the underlying quadratic function. This property isn't true of all probability distributions (for example, the mean may be between two peaks at a relatively low probability point); it's another one of the properties that makes manipulating Gaussians nice.

3.1 A detour into quadratic functions and geometry

This section can be skipped on first read since it's somewhat technical. It expands on the geometry of the Gaussian by analyzing the underlying quadratic function in the exponent; a strong understanding of how the geometry of multidimensional quadratics relates to the mean, covariance, and inferential properties of the associated Gaussian builds intuition for their use as building blocks in a number of problems. That said, this material isn't absolutely critical to be able to follow the Kalman filter derivation in Section 4.

3.1.1 Stretched isometric quadratics

Lets examine the probability density function of a generic multivariate Gaussian distribution with mean μ and covariance Σ:

$$p(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}, \tag{4}$$

Fundamental to this density is the quadratic function

$$f(x; \mu, \Sigma) = \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)$$

in the exponent. When f is large, p is small, and vice versa. And more, the quicker f increases, the quicker p decreases—symmetries in the shape of f are reflected in the shape of p. Below, we'll see that each f is really just a stretched (or squished) and shifted version of a simple uniformly symmetric "canonical" quadratic function $c(u) = \frac{1}{2}\|u\|^2$. And more than that, this stretching and squishing happens along well-defined orthogonal axes defined by the properties of the covariance matrix Σ. Specifically, the Eigenvectors of Σ will tell us in which directions the stretching occurs, and the (square roots of) the Eigenvalues will tell us how strongly the space is stretched or squished along those dimensions.

To start we need to review what may be viewed as the canonical coordinate system of a positive definite matrix. For those unfamiliar with Eigenvalues and/or the Eigen-decomposition of positive definite matrices, I've included some background material in Appendix A.

Denote the Eigenvalues and Eigenvectors of Σ by $\{\sigma_i^2\}_{i=1}^n$ and $\{e_i\}_{i=1}^n$, respectively. Using the expansion of Σ in terms of its Eigenvectors and Eigenvalues

$$\Sigma = \sum_i \sigma_i^2 \; e_i e_i^T$$

(see Appendix A), and the properties

$$\left(\sum_i \sigma_i^2 \; e_i e_i^T\right)^{-1} = \sum_i \frac{1}{\sigma_i^2} \; e_i e_i^T \qquad \text{and} \qquad \left(\sum_i \sigma_i \; e_i e_i^T\right)^2 = \sum_i \sigma_i^2 \; e_i e_i^T,$$

we can rewrite the quadratic function as

$$\begin{aligned} f(x) &= \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \\ &= \frac{1}{2}\left[\left(\sum_{i=1} \frac{1}{\sigma_i} e_i e_i^T\right)(x-\mu)\right]^T \left[\left(\sum_{i=1} \frac{1}{\sigma_i} e_i e_i^T\right)(x-\mu)\right] \\ &= \frac{1}{2} u^T u, \end{aligned}$$

where $u = \left(\sum_{i=1} \frac{1}{\sigma_i} e_i e_i^T\right)(x-\mu)$. Thus, in terms of u, the quadratic function is just a very simple spherically symmetric function $\frac{1}{2}\|u\|^2$. In order to get back to the original function, we need to apply the inverse transform

$$x = \left(\sum_{i=1} \sigma_i \; e_i e_i^T\right) u + \mu, \tag{5}$$

which, from our analysis in Section 3.1, is just a stretching of the space along the directions e_i by factors σ_i and a shift by μ.

Looking back at the Gaussian density function of Equation 4, we can now examine how this transformation affects the density. Consider a very simple *isometric* Gaussian distribution with mean 0 and covariance I (the identity matrix)[6]

$$\begin{aligned} p(u; 0, I) &= \frac{1}{\sqrt{(2\pi)^n |I|}} \exp\left\{-\frac{1}{2} u^T I u\right\} \\ &= \frac{1}{(\sqrt{2\pi})^n} \exp\left\{-\frac{1}{2}\sum_{i=1}^n u_i^2\right\} \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{u_i^2}{2}}. \end{aligned}$$

[6] By *isometric* we mean that the Gaussian is the same in all directions of the space. Geometrically, there's no way to distinguish one direction from any other direction.

That final expression shows that the zero-centered multi-dimensional Gaussian with (isometric) covariance I is actually just a product of a bunch of independent one-dimensional normal distributions.

This isometric density, though, is just a Gaussian built from the simplest quadratic function we considered above. To get any arbitrary Gaussian $\mathcal{N}(\mu, \Sigma)$, we can just apply the stretching transformation of Equation 5. Intuitively, this transformation stretches the space as defined by the Eigenspace of Σ and then shift it by μ.

Formalizing this intuition is a straightforward application of our quadratic function manipulation procedure outlined above. If we translate u into stretched coordinates via Equation 5, that corresponds to stretching the quadratic in the exponent from the isometric $\frac{1}{2}u^T u$ to the elongated $\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)$. Trivially so, the resulting density is now the density for a normal distribution with mean μ and covariance Σ.

The next section shows that we can get this same result by applying this transformation directly to the random variable $U \sim \mathcal{N}(0, I)$.

3.1.2 Transformed Gaussian random variables.

One straightforward method for calculating how linear transformations affect Gaussian distributions is by manipulating the random variables directly. Since a Gaussian distribution is fully specified by its mean and covariance, we need only compute these two *statistics* using the formulas discussed in Section 2.1 to find how the transformation affects the distribution. This technique significantly simplifies our derivation of the Kalman filter in Section 4.

Starting from a random variable $U \sim \mathcal{N}(0, I)$, as we showed in Sections 2.2 and 2.3, transforming the random variable as

$$X = \underbrace{\left(\sum_{i=1}^{n} \sigma_i \; e_i e_i^T\right)}_{\Sigma^{\frac{1}{2}}} U + \mu$$

gives us a new random variable with mean μ and covariance $(\Sigma^{\frac{1}{2}})^T I \Sigma^{\frac{1}{2}} = \Sigma$. We already knew this based on the above analysis of the density, and more than that, the above analysis showed that this new random variable is again a Gaussian distribution, but we now understand that all Gaussian random variables are simply affine transformations of the random variable U. This observation makes it clear that each Gaussian is actually just a stretched and shifted version of the simple canonical Gaussian represented by U, and it's an way to reason about Gaussian random variables and calculate their parameters (the mean and covariance).

4 The Kalman filter

Now that we understand how to manipulate Gaussian random variables, deriving the Kalman filter is straightforward. This section assumes a background in Bayesian filtering and focuses primarily on the calculation of explicit update formulas under Gaussian assumptions.

To formulate the problem, let $X_t \sim \mathcal{N}(\widehat{x}_t, \Sigma_t)$ be a normally distributed random variable denoting our uncertainty about our state currently at time t. At this time step, we're going to take an action a_t to take us into the next time step, and we assume our probabilistic motion model takes the convenient affine form

$$X_{t+1} = AX_t + a_t + \epsilon_t,$$

where $\epsilon_t \sim \mathcal{N}(0, Q)$ denotes Gaussian noise injected into the system. In the next time step, we'll also make an observation y_{t+1} with distribution dependent on the state predicted by our motion model

$$Y_{t+1} = BX_{t+1} + v_t + \delta_t.$$

Here v_t is just a non-random offset vector, and $\delta_t \sim \mathcal{N}(0, R)$ is again Gaussian noise injected into the system. The Kalman filter updates simply express the posterior Gaussian parameters (the mean and covariance) of X_{t+1} given Y_{t+1} as a function of the previous time step's parameters and the coefficients of the motion and observation models.

We can find these update rules by first computing the joint distribution over both X_{t+1} and Y_{t+1} and then invoking the conditioning identities in Equations 9 and 10. This computation involves calculating the mean and covariance of the two multivariate random variables X_{t+1} and Y_{t+1} as well as the covariance between X_{t+1} and Y_{t+1}, and we can make all of these calculations by simply invoking the covariance definition detailed in Section 2.1.

The mean of both X_{t+1} and Y_{t+1} are straightforward to calculate

$$\begin{aligned}\widehat{x}'_{t+1} &= A\widehat{x}_t + a_t \\ \widehat{y}_{t+1} &= B\widehat{x}_{t+1} + v_t.\end{aligned}$$

We use the $\widehat{x}'_{t+1}$ to denote the mean of X_{t+1} rather than simply $\widehat{x}_{t+1}$ because it represents the mean of our state after taking the action but before accounting for the observation. $\widehat{x}_{t+1}$ is reserved for the mean of the conditional distribution over next states given the observation.

To find the covariance of X_{t+1}, we calculate

$$\begin{aligned}E[(X_{t+1} - \widehat{x}'_{t+1})(X_{t+1} - \widehat{x}'_{t+1})^T] &= E[(A(X_t - \widehat{x}_t) + \epsilon_t)(A(X_t - \widehat{x}_t) + \epsilon_t)^T] \\ &= A\Big(E[(X_t - \widehat{x}_t)(X_t - \widehat{x}_t)^T]\Big)A^T + E[\epsilon_t\epsilon_t^T] \\ &= A\Sigma_t A^T + Q = \Sigma'_{t+1}.\end{aligned}$$

Note that cross terms left out of this expression, such as $E[A(X_t - \widehat{x}_t)\epsilon_t]$, have expectation zero since the associated random variables are independent of one another. Similarly, the covariance of Y_{t+1} is

$$\begin{aligned}
&E[(Y_{t+1} - \widehat{y}_{t+1})(Y_{t+1} - \widehat{y}_{t+1})^T] \\
&\quad = E[(B(X_{t+1} - \widehat{x}'_{t+1}) + \delta_t)(B(X_{t+1} - \widehat{x}'_{t+1}) + \delta_t)^T] \\
&\quad = B\left(E[(X_{t+1} - \widehat{x}'_{t+1})(X_{t+1} - \widehat{x}'_{t+1})^T] \right) B^T + E[\delta_t \delta_t^T] \\
&\quad = B\Sigma'_{t+1}B^T + R.
\end{aligned}$$

Finally, the covariance between X_{t+1} and Y_{t+1} is

$$\begin{aligned}
E[(X_{t+1} - \widehat{x}'_{t+1})(Y_{t+1} - \widehat{y}_{t+1})^T] &= \left(E[(X_{t+1} - \widehat{x}'_{t+1})(X_{t+1} - \widehat{x}'_{t+1})^T] \right) B^T \\
&= \Sigma'_{t+1}B^T
\end{aligned}$$

All together, this gives us a joint distribution with the form

$$p(X_{t+1}, Y_{t+1}) \sim \mathcal{N}\left(\begin{bmatrix} A\widehat{x}_t + a_t \\ B\widehat{x}'_{t+1} + v_t \end{bmatrix}, \begin{bmatrix} \Sigma'_{t+1} & \Sigma'_{t+1}B^T \\ B\Sigma'_{t+1} & B\Sigma'_{t+1}B^T + R \end{bmatrix} \right).$$

Now we can leverage the Gaussian conditional formulas in Equations 9 and 10 to directly write down the Kalman filter update

$$\begin{aligned}
\widehat{x}'_{t+1} &= A\widehat{x}_t + a_t \\
\Sigma'_{t+1} &= A\Sigma_t A^T + Q \\
x_{new} &= \widehat{x}'_{t+1} + \Sigma'_{t+1}B^T \left(B\Sigma'_{t+1}B^T + R\right)^{-1} (y_{t+1} - \widehat{y}_{t+1}) \\
\Sigma_{t+1} &= \Sigma'_{t+1} - \underbrace{\Sigma'_{t+1}B^T \left(B\Sigma'_{t+1}B^T + R\right)^{-1}}_{K} B\ \Sigma'_{t+1},
\end{aligned}$$

where, as indicated, we often substitute $K = \Sigma'_{t+1}B^T \left(B\Sigma'_{t+1}B^T + R\right)^{-1}$ to simplify the expression.

5 Gaussian inference and optimization

[This section has not yet been written.]

A Appendix: Eigenvectors, Eigenvalues, and the Eigen-decomposition

For those not familiar with Eigenvectors and Eigenvalues, here's a quick review of their properties. An Eigenvector e of a matrix Σ is a vector who's direction doesn't change when transformed by Σ. It's likely stretched by some factor λ, but its direction doesn't change: $\Sigma e = \lambda e$. In general, λ can be negative, but when Σ is a covariance matrix we'll find below that λ is always be positive.

A.1 A quick note on positive definiteness and positive Eigenvalues

A positive definite matrix A is one for which $v^T Av > 0$ for all v as long as v isn't the zero vector (in which case $0^T A0 = 0$). An example of a positive definite matrix is the covariance matrix of a multivariate random variable X. As we defined in Section 2.3, the covariance matrix is

$$\Sigma = E[(X - \mu)(X - \mu)^T],$$

which means for any v,

$$\begin{aligned} v^T \Sigma v &= v^T E[(X - \mu)(X - \mu)^T] v = E[(v^T (X - \mu))^2] \\ &= \int (v^T (x - \mu))^2 \; p(x) dx > 0. \end{aligned}$$

The last inequality is true because because each function value inside the integral is non-negative, and there are some (formally, "measurable") regions of the domain for which it's strictly positive. Thus, the covariance matrix is always positive definite.

If any of the Eigenvalues λ_k of Σ are negative or zero, we'd have some vector (for instance e_k) for which

$$e_k^T \Sigma e_k = \lambda_k e_k^T e_k = \lambda_k \leq 0,$$

which would contradict the positive definiteness property. All Eigenvalues must, therefore, be strictly positive since Σ is positive definite. When discussing Eigenvalues in the context of covariance matrices, we often write $\lambda_i = \sigma_i^2$ to simultaneously emphasize it's positively and its relationship to squared standard deviations (variances).

A full discussion of Eigen-decompositions (diagonalizations) of symmetric positive definite matrices is beyond the scope of this document, so we'll just take it to be true that all symmetric positive definite matrices have a full set of n Eigenvectors and Eigenvalues, and that set of Eigenvectors form an orthonormal basis $\mathcal{B} = \{e_i\}_{i=1}^n$ for the n-dimensional domain (and range) spaces of the matrix.

The behavior of a matrix is entirely defined by it's behavior on elements of a basis,[7] so since the matrix $\sum_{i=1}^n \sigma_i^2 \left(e_i e_i^T\right)$ behaves like Σ on each basis element e_i (specifically, $\sum_{i=1}^n \sigma_i^2 \left(e_i e_i^T\right) e_k = \sigma_k^2 e_k$ since e_i is orthogonal to e_k when $i \neq k$, and $e_k^T e_k = 1$), we can conclude that this expression is a valid expansion of the covariance matrix:

$$\Sigma = \sum_{i=1}^n \sigma_i^2 \left(e_i e_i^T\right). \tag{6}$$

[7] For any matrix A and basis $\{v_i\}_{i=1}^n$, the "action" of A on a vector is $Ax = A \sum_i \alpha_i v_i = \sum_i \alpha_i A v_i$, where α_i are the unique coefficients of expansion in terms of the given basis.

This decomposition in terms of the Eigenvalues and Eigenvalues of Σ, sometimes termed the Eigen-decomposition or diagonalization of the matrix,[8] emphasizes the geometry of the linear transformation defined by Σ.

A.2 A window into the geometry of Σ

The decomposition in Equation 6 is a window into the geometrical properties of Σ, and is particularly important in understanding Σ's role in defining the geometry of quadratic functions.

Since actually the square root of Σ is most relevant to the discussion in Section 3.1.1, we'll consider specifically here the matrix

$$\Sigma^{\frac{1}{2}} = \sum_{i=1}^{n} \sigma_i \left(e_i e_i^T \right), \tag{7}$$

although all matrices formed by scaled sums of the outer product terms $e_i e_i^T$ fit the discussion below.

Each Eigenvector e_i is both normalized and orthogonal to all other Eigenvectors e_j for $j \neq i$, which means that each term $\sigma_i(e_i e_i^T)$ of the square-root expansion in Equation 7 is proportional to the one-dimensional projection matrix $e_i e_i^T$ which projects a vector v onto the Eigenvector: $[e_i e_i^T]\ v = (e_i^T v) e_i$. These terms therefore explicitly stretch the vector in the directions e_i by factors σ_i. Since each term acts orthogonally to all other terms none of these stretches (or contractions in the case $\sigma_i < 1$) affects any of the others. This matrix literally just stretches or contracts the space in the directions e_i by factors σ_i.

This geometry is made explicit with a little more math. Since the set of vectors $\{e_i\}_{i=1}^n$ forms a basis for the space, we can expand any vector v as $v = \sum_{i=1}^n \alpha_i e_i$ as we did above.[9] The action of $\Sigma^{\frac{1}{2}}$ as an operator can then be analyzed as

$$\begin{aligned}
\Sigma^{\frac{1}{2}} v &= \left(\sum_{i=1}^{n} \sigma_i\ e_i e_i^T \right) \left(\sum_{j=1}^{n} \alpha_j e_j \right) \\
&= \sum_{i=1}^{n} \sigma_i\ e_i \sum_{j=1}^{n} \alpha_j e_i^T e_j \\
&= \sum_{i=1}^{n} \sigma_i \alpha_i\ e_i,
\end{aligned}$$

since each term $e_i^T e_j$ is zero when $i \neq j$ and 1 when $i = j$. This bit of algebra explicitly says that multiplying by this matrix $\Sigma^{\frac{1}{2}}$ stretches the vector by a factor of σ_i in direction e_i in the way we discussed intuitively above.

[8] Written in matrix form, this decomposition becomes $\Sigma = UDU^T$, where U is an orthogonal matrix whose columns consist of the Eigenvectors of Σ and D is a diagonal matrix whose diagonal contains the corresponding Eigenvalues.

[9] Since the Eigenvectors are normalized and orthogonal to one another, in this case each $\alpha_i = e_i^T v$.

B Understanding general covariance through the SVD

The next section takes the above analysis one step further, using the SVD to analyze a general covariance matrix between two different multivariate random variables.

B.1 Geometric perspectives on the SVD

The Fundamental Theorem of Linear Algebra states that every $m \times n$ matrix A can be decomposed in the form $A = USV^T$, where U and V are both square orthogonal matrices of dimensions m^2 and n^2, respectively, and S is a diagonal matrix containing what are known as the *singular values* $\sigma_i{}_{i=1}^{m}$ of A. In this section, we'll generally consider the case $m < n$, but other cases can can be analyzed analogously.

The singular values are closely related to Eigenvalues—indeed, many proofs or the Fundamental Theorem of Linear Algebra proceed[10] constructively by first forming the Eigen-decomposition of the symmetric positive definite matrix AA^T. When A isn't square ($m \neq n$), S mitigates the difference in dimension with added zeros for buffering. For instance, for our setting where $m < n$, V is going to be larger than U, so we typically write

$$A = U \begin{bmatrix} \widetilde{S} & 0 \end{bmatrix} \begin{bmatrix} V_{/\!/}^T \\ V_\perp^T \end{bmatrix}, \tag{8}$$

where $\widetilde{S}$ is a square $n \times n$ diagonal matrix. This explicit notation illuminates the role of the separate blocks of $V = [V_{/\!/}\ V_\perp]$. When multiplied through, the section of zeros in S obliterates the submatrix $V_\perp$, so really, we could have written it simply as $A = U\widetilde{S}V_{/\!/}^T$. Writing it as we do, though, in Equation 8, emphasizes that any components of a vector x lying within the span of $V_\perp$ (or, equivalently, we could say any components of x perpendicular to $V_{/\!/}$ since V is orthogonal) vanish when we multiply by A. The following calculation illustrates this phenomenon:

$$\begin{aligned} Ax &= U \begin{bmatrix} \widetilde{S} & 0 \end{bmatrix} \begin{bmatrix} V_{/\!/}^T \\ V_\perp^T \end{bmatrix} x = U \begin{bmatrix} \widetilde{S} & 0 \end{bmatrix} \begin{bmatrix} V_{/\!/}^T x \\ V_\perp^T x \end{bmatrix} \\ &= U\widetilde{S}V_{/\!/}^T x. \end{aligned}$$

[10] To sketch the idea, we can decompose $AA^T = U\widetilde{S}^2U^T$ where the columns of U contain the Eigenvectors and $\widetilde{S}^2$ is a diagonal matrix whose entries are the Eigenvectors σ_i^2. Since U and $\widetilde{S}$ are invertible (the former is orthogonal and the latter is diagonal), we can write $A = U\widetilde{S}B^T$ for some B, as of yet unknown. However, $AA^T = U\widetilde{S}^2U^T$, so it must be that $B^TB = I$. The columns of B are, therefore, mutually orthogonal and normalized, so we can write $B = V_{/\!/}$. The space orthogonal to $V_{/\!/}$ doesn't matter (as we'll see below since it vanishes upon multiplication), so we can just fill the rest of V with any orthogonal basis of the space orthogonal to $V_{/\!/}$ and call it $V_\perp$.

In other words, when we expand $x = V_{/\!/}x + V_\perp x$ in terms of its components that lie within the orthogonal subspaces spanned by $V_{/\!/}$ and $V_\perp$, we see immediately that the component $V_\perp x$ vanishes entirely from the final expression because it's eaten by the section of zero within $S = [\widetilde{S}\ 0]$.

In linear algebra, we have names for the space spanned by $V_{/\!/}$ and $V_\perp$. The first subspace, denoted $\mathbf{span}(V_{/\!/})$, is called the *row space* of A since it spans the same space as the rows of A itself.[11] The second, $\mathbf{span}(V_\perp)$ is called the *nullspace*. Any vector within the nullspace (or any vector component lying within the nullspace) has no effect during transformation by A—it's annihilated by the zero padding of S. That nullspace property results in a fundamental way from the structure of A. If the rows of A are all independent (for $m < n$), then since we can always perform the SVD, the above explicit decomposition given in Equation 8 shows that there's no room for the columns of $V_\perp$ to have any effect on the transformation. Finally, $\mathbf{span}(U)$ is called the *column space* of A since, using a similar argument for analyzing $\mathbf{span}(V_{/\!/})$, we can show that the columns of A and the columns of U span the same spaces. Moreover, the respective columns of all three of these matrices, $V_{/\!/}$, $V_\perp$, and U, form orthonormal bases for their respective spaces.

Now if we write this decomposition in a way analogous to the Eigen-decomposition of Equation 6, we can gain insight into the behavior of matrix:

$$A = \sum_{i=1}^{m} \sigma_i \ u_i v_i^T,$$

where u_i is the ith column of U and v_i is the ith column of $V_{/\!/}$. This expression shows us that the matrix is essentially identifying the ith basis vector v_i of the subspace spanned by $V_{/\!/}$ with the ith basis vector u_i of the subspace spanned by U. Multiplying a vector x by A, therefore, computes the following:

$$\begin{aligned} Ax &= \sum_{i=1}^{m} \Big(\ \sigma_i (v_i^T x) \ \Big) \ u_i = \sum_{i=1}^{m} \left(\ \sigma_i (v_i^T \sum_j \alpha_j v_j) \ \right) \ u_i \\ &= \sum_{i=1}^{m} \Big(\ \sigma_i \alpha_i \ \Big) \ u_i \end{aligned}$$

where we've used the expansion of x in terms of the basis $\{v_i\}_{i=1}^n$. In other words, the matrix looks at the component of x along a basis vector v_i of the domain, stretches it by a factor σ_i, and assigns it as the coefficient to the basis

[11] That's straightforward to see remembering that U and $\widetilde{S}$ are both invertible—there's a one-to-one mapping between the coefficients that reconstruct a vector in the row-span of A the coefficients that reconstruct the vector in the row-span of $V_{/\!/}^T$. More explicitly, if $v^T = \beta^T A$ for some coefficients β, then the coefficients $\widetilde{\beta}^T = \beta^T U \widetilde{S}$ are coefficients such that $\widetilde{\beta} V_{/\!/}^T = v^T$. Likewise, for any vector u in the span of $V_{/\!/}$, there are coefficients that construct that vector as $u^T = \widetilde{\alpha}^T V_\perp^T$. Then since $\widetilde{\alpha}^T V_{/\!/}^T = \widetilde{\alpha}^T (\widetilde{S}^{-1} U^T) U \widetilde{S} V_{/\!/}^T = \alpha^T U \widetilde{S} V_{/\!/}^T$, where $\alpha = U \widetilde{S}^{-1} \widetilde{\alpha}$, we can say that u^T is in the row-span of the $A = U \widetilde{S} V_{/\!/}^T$.

vector u_i in the range space! You can think of this operation as taking the space $\mathbf{span}(V_{/\!/})$, which is a linear subspace of the domain, stretching it in the directions v_i by the factors σ_i, and fusing it with the range space $\mathbf{span}(U)$ so that the bases align. Every matrix with $m < n$ is an identification between a stretched version of an m-dimensional linear subspace of the domain and the m-dimensional vector space of the range.

B.2 Relevance to general non-symmetric covariance matrices

What does this mean for Covariance matrices? So far we've discussed the geometry of symmetric positive definite covariance matrices $\text{Cov}(X)$ that express the inter-covariances between the individual variables of a single multivariate random variable X. As we saw in Section 2.3, though, covariance matrices often more generally express the covariance *between* the individual variables of one multivariate random variable X and another Y. In this setting, denoting the mean of X by μ_x and the mean of Y by μ_y, the covariance matrix is $\Sigma_{xy} = E[(X-\mu_x)(Y-\mu_y)^T]$, which is an $m \times n$ non-square matrix where m and n are determined by the dimensionalities of X and Y, respectively. Analyzing the geometry of this covariance matrix requires we use the more general Singular Value Decomposition introduced above. This analysis gives us the tools necessary to understand the geometry of the Conditional Gaussian formula presented below in Section C at an intuitive level.

The covariance matrix in this setting tells us how various directions in the space of X vary with directions in the space of Y. It's easiest to think the geometry in terms of the trends seen in a collection of samples $\{(x_j, y_j)\}_{j=1}^N$ taken from the joint distribution over both random variables $p(X, Y)$. Then we can ask how a specific component of x_j in the direction of some vector v in the domain space[12] $\mathcal{X}$ tends to vary with the associated component of y_j along a vector u in the range space $\mathcal{Y}$. The left and right singular vectors $\{u_i\}_{i=1}^m$ and $\{v_i\}_{i=1}^m$, in conjunction with the singular values $\{\sigma_i\}_{i=1}^m$, hand us this information directly. When σ_i is large, components of the vectors y_j along v_i will tend be large when components of x_j along u_i are large. Likewise, when σ_i is small in magnitude, components of y_j along v_i won't really correspond in any consistent way with components of x_j along u_i. Finally, in this general case, we can have negative σ_i, which means that components of y_j along v_i will tend to vary negatively with components of x_j along u_i (when one's largely positive, the other will be largely negative).

This analysis tells us that although the multivariate random variable X consists of m random variables and the multivariate random variable Y consists of n random variables, where $m < n$, there are actually only m *principle* one-dimensional random variable pairs $(U_i, V_i)_{i=1}^m$, each representing components of X and Y along u_i and v_i, that co-vary with each other as given by σ_i. These principle random variables are entirely independent of all other pairs of random

[12] Specifically, that component is $\widehat{v}^T x_j$, where $\widehat{v}$ is the normalized version of v.

variables in the sense that V_i has zero covariance with all other random variables V_j and U_j when $i \neq j$. Every pair of multivariate random variables X and Y fundamentally co-vary independently along fixed corresponding axes given by the singular vectors. The degree to which they co-vary is given by the singular values.

C Conditional Gaussians

This section recites without proof the formula for finding the mean and covariance of a Gaussian of the form $p(x|y)$ when you know the explicit form of the joint distribution $p(x, y)$.

We can write the joint distribution blockwise as

$$\begin{aligned} p(x,y) &= \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}\right) \\ &\propto \exp\left\{-\frac{1}{2}\begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix}^T \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}^{-1} \begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix}\right\} \end{aligned}$$

Then the conditional distribution $p(x|y)$ is a Gaussian as well with mean and covariance

$$\mu_{x|y} = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y) \tag{9}$$

$$\Sigma_{x|y} = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}. \tag{10}$$

Note that in these equations, since the covariance matrix is symmetric, $\Sigma_{xy} = \Sigma_{yx}^T$.

We can understand these equations intuitively by leveraging the geometric analysis outlined above. Consider Equation 9 defining the conditional mean, for instance, in the context of filtering, where semantically x represents a state variable and y represents an observation (strictly, the distribution over y would be the probability of y *given* a hypothesized state). Before making an observation, we predict what the observation would likely be μ_y given what we've predicted our state to be. That prediction's usually somewhat wrong; the more wrong it is, the more we should correct the state prediction. The matrix $\Sigma_{xy}\Sigma_{yy}^{-1}$ explicitly transforms the error in observation space (between our predicted observation μ_y and the actual observation y) into a corrective perturbation to the predicted state μ_x. The more wrong our predicted observation was, the larger we'd expect that perturbation to be; in a sense, large differences between predicted and actual observations contain more information than smaller differences.

But more than that, based on our observation model, there are some directions in observation space that are more likely to have variations than others. The profile of these expected observation variances is handed to us by the observation model's covariance matrix Σ_{yy}. Directions under which there's significant variance don't give us much information since we already expect to see that sort of variation given our current state estimate. But if we receive a particular observation y that varies from μ_y significantly along an unusual direction (one under

which our observation model says there's typically small variance), then that observation's unique and it potentially tells us a lot about how the actual state might differ from the state we've estimated prior to making the observation.

Multiplying $\delta y = y - \mu_y$ by the inverse of Σ_{yy} means that we shrink components δy in directions of high variance, and amplify it in directions of low variance. This operation transforms δy based on how much information each component provides.

The second matrix multiplication by Σ_{xy} then transforms this *information scaled* observation offset vector $\delta\widetilde{y} = \Sigma_{yy}^{-1}\delta y$ into the state-space to directly perturb the expected state vector μ_x. As we saw above in Section B, this covariance matrix matches orthogonal directions v_i in observation space to corresponding orthogonal directions u_i in the state space. Variation from the mean in observation space along a direction v_i co-varies with variation from the mean in state space along direction u_i, but variations in those directions are entirely independent of variations along any of the other directions u_j or v_j $(i \neq j)$. Thus, if our information scaled observation offset $\delta\widetilde{y}$ has significant components along directions that tend to co-vary strongly with the state vector, that observation tells us significant additional information about the state. The matrix Σ_{xy} then literally translates $\delta\widetilde{y}$ to a significant perturbation to the expected state μ_x along that direction.

On the other hand, if much of information scaled observation offset $\delta\widetilde{y}$ lies along directions that don't co-vary significantly with the state, this observation doesn't really tell us much about the state. That's reflected in the multiplication by Σ_{xy}, which will annihilate much of the vector along those directions. In the extreme case, the offset could lie entirely in the nullspace of Σ_{xy} and thereby hold absolutely zero information about the state. Even if the observation y could potentially tell us a lot of information (e.g. the information scaled $\Sigma_{yy}^{-1}(y - \mu_y)$ is large in magnitude indicating that the observation is unique and unexpected, full of *potential* information about the state), if that observation doesn't really correlate well with the state (the covariance between the observation variable along that direction and the state is close to zero), then that observation still doesn't tell us much about the state.

In order for the observation to be informative, it needs to be both 1. unique and unexpected (differs significantly from what we expected based on our motion model state prediction), and 2. well-correlated with the state. Equation 9 reflects both of these criteria through the transformation $\Sigma_{xy}\Sigma_{yy}^{-1}$.

A similar set of arguments can be made to understand how the act of making an observation can resolve some of the uncertainty. I won't go into detail about this one beyond to say that, examining Equation 10, we see that we start with Σ_{xx} and subtract from that a term that represents the inverse of Σ_{yy} as viewed from within the state space (sandwiched between operators Σ_{xy} and Σ_{yx} which represent how variation within Y corresponds to variation with X). This says that small variance directions in Y correspond to large reductions in variance after conditioning since they're highly informative.