Week 1
Discrete time Gaussian Markov processes

Jonathan Goodman

September 10, 2012

1 Introduction to Stochastic Calculus

These are lecture notes for the class *Stochastic Calculus* offered at the Courant Institute in the Fall Semester of 2012. It is a graduate level class. Students should have a solid background in probability and linear algebra. The topic selection is guided in part by the needs of our MS program in Mathematics in Finance. But it is not focused entirely on the Black Scholes theory of derivative pricing. I hope that the main ideas are easier to understand in general with a variety of applications and examples. I also hope that the class is useful to engineers, scientists, economists, and applied mathematicians outside the world of finance.

The term *stochastic calculus* refers to a family of mathematical methods for studying dynamics with randomness. *Stochastic* by itself means *random*, and it implies dynamics, as in *stochastic process.* The term *calculus* by itself has two related meanings. One is a system of methods for calculating things, as in the *calculus of pseudo-differential operators* or the *umbral calculus.*[1] The tools of stochastic calculus include the *backward equations* and *forward equations*, which allow us to calculate the time evolution of expected values and probability distributions for stochastic processes. In simple cases these are matrix equations. In more sophisticated cases they are partial differential equations of diffusion type.

The other sense of *calculus* is the study of what happens when $\Delta t \to 0$. In this limit, finite differences go to derivatives and sums go to integrals. Calculus in this sense is short for *differential calculus* and *integral calculus*,[2] which refers to the simple rules for calculating derivatives and integrals – the product rule, the fundamental theorem of calculus, and so on. The operations of calculus, integration and differentiation, are harder to justify than the operations of algebra. But the formulas often are simpler and more useful: integrals can be easier than sums.

[1] Just name-dropping. These are not part of our Stochastic Calculus class.

[2] Richard Courant, the founder of the Courant Institute, wrote a two volume textbook *Differential and Integral Calculus.* (originally *Vorlesungen über Differential und Integralrechnung*).

Differential and integral calculus is good for modeling as well as for calculation. We understand the dynamics of a system by asking how the system changes in a small interval of time. The mathematical model can be a system of differential equations. We predict the behavior of the system by solving the differential equations, either analytically or computationally. Examples of this include rate equations of physical chemistry and the laws of Newtonian dynamics.

The stochastic calculus in this sense is the *Ito calculus.* The extra *Ito term* makes the Ito calculus more complicated than ordinary calculus. There is no Ito term in ordinary calculus because the *quadratic variation* is zero. The Ito calculus also is a framework for modeling. A stochastic process may be described by giving an Ito stochastic differential equation, an *SDE.* There are relatively simple rules for deriving this SDE from basic information about the short time behavior of the process. This is analogous to writing an ordinary differential equation (*ODE*) to describe the evolution of a system that is not random. If you can describe the behavior of the system over very short time intervals, you can write the ODE. If you can write the ODE, there is an array of analytical and computational methods that help you figure out how the system behaves.

This course starts with two simple kinds of stochastic processes that may be described by basic methods of probability. This week we cover linear Gaussian recurrence relations. These are used throughout science and economics as the simplest class of models of stochastic dynamics. Almost everything about linear Gaussian processes is determined by matrices and linear algebra. Next week we discuss another class of random processes described by matrices, finite state space Markov chains. Week 3 begins the transition to continuous time with continuous time versions of the Gaussian processes discussed this week. The simplest of these processes is Brownian motion, which is the central construct that drives most of the Ito calculus.

After than comes the technical core of the course, the Ito integral, Ito's lemma, and general diffusion processes. We will see how to associate partial differential equations to diffusion processes and how to find approximate numerical solutions.

It is impractical to do all this in a mathematically rigorous way in just one semester. This class will indicate some of the main ideas of the mathematically rigorous theory, but we will not discuss them thoroughly. Experience shows that careful people can Ito calculus more or less correctly without being able to recite the formal proofs. Indeed, the ordinary calculus of Newton and Leibnitz is used daily by scientists and engineers around the world, most of whom would be unable to give a mathematically correct definition of the derivative.

Computing is an integral part of this class as it is an integral part of applied mathematics. The theorems and formulas of stochastic calculus are easier to understand when you see them in action in computations of specific examples. More importantly, in the practice of stochastic calculus, it is very rare that a problem gets solved without some computation. A training class like this one should include all aspects of the subject, not just those in use before computers were invented.

2 Introduction to the material for the week

The topic this week is linear recurrence relations with Gaussian noise. A *linear recurrence relation* is an equation of the form

$$X_{n+1} \;=\; AX_n \,+\, V_n \;. \qquad (1) \quad \boxed{\texttt{lrr}}$$

The *state vector* at time n is a column vector with d components:

$$X_n \;=\; \begin{pmatrix} X_{n,1} \\ \vdots \\ X_{n,d} \end{pmatrix} \;\in\; \mathbb{R}^d \;.$$

The *innovation*, or *noise vector*, or *random forcing*, is the d component V_n. The forcing vectors are *i.i.d.,* which stands for *independent* and *identically distributed.* The model is defined by the matrix A and the probability distribution of the forcing. The model does not change with time because A and the distribution of the V_n are the same for each n. The recurrence relation is *Gaussian* if the noise vectors V_n are Gaussian.

This is a simple model of the evolution of a system that is somewhat predictable, but not entirely. The d components $X_{n,1}, \ldots, X_{n,d}$ represent the state of the system at time n. The deterministic part of the dynamics is $X_{n+1} = AX_n$. This says that the components at the next time period are linear functions of the components in the current period. The term V_n represents the random influences at time n. In this model, everything about time n relevant to predicting time $n+1$ is contained in X_n. Therefore, the noise at time n, which is V_n, is completely independent is anything we have seen before.

In statistics, a model of the form $Y = AX + V$ is a *linear regression* model (though usually with $\mathrm{E}[V] = 0$ and V called ϵ). In it, a family of variables Y_i are predicted using linear functions of a different family of variables X_j. The noise components V_k model the extent to which the Y variables cannot be predicted by the X variables. The model (1) is called *autoregressive*, or *AR*, because values of variables X_j are predicted by values of the same variables at the previous time.

It is possible to understand the behavior of linear a Gaussian AR model in great detail (technical detail, you must assume it starts with X_0 that is Gaussian). All the subsequent X_n are multivariate normal. There are simple matrix recurrence relations that determine their means and covariance matrices. These recurrence relations important in themselves, and they are our first example of an ongoing theme for the course, backward and forward equations.

This material makes a good start for the class for several reasons. For one thing, it gives us an excuse to review multivariate Gaussian random variables. Also, we have a simple context in which to talk about *path space.* In this case, a *path* is a sequence of consecutive states, which we write as

$$X_{[n_1:n_2]} \;=\; \left(X_{n_1}, X_{n_1+1}, \ldots, X_{n_2}\right) \,. \qquad (2) \quad \boxed{\texttt{p}}$$

The notation $[n_1 : n_2]$ comes from two sources. Several programming languages use similar notation to denote a sequence of consecutive integers: $[n_1 : n_2] = \{n_1, n_1+1, \ldots, n_2\}$. In mathematics, $[n_1, n_2]$ refers to the closed interval containing all real numbers between n_1 and n_2, including n_1 and n_2. We write $[n_1 \colon n_2]$ to denote just the integers in that interval.

The path is an object in a big vector space. Each of the X_n has d components. The number of integers n in the set $[n_1 : n_2]$ is $n_2 - n_1 + 1$. Altogether, the path $X_{[n_1:n_2]}$ has $d(n_2 - n_1 + 1)$ components. Therefore $X_{[n_1:n_2]}$ can be viewed as a point in the path space $\mathbb{R}^{d(n_2-n_1+1)}$. As such it is Gaussian. Its distribution is completely determined by its mean and covariance matrix. The mean of $X_{[n_1:n_2]}$ is determined by the means of the individual X_n. The covariance matrix of $X_{[n_1:n_2]}$ has dimension $d(n_2 - n_1 + 1) \times d(n_2 - n_1 + 1)$. Some of its entries give the variances and covariances of the components of X_n. Others are the covariances of compenents $X_{n,j}$ with $X_{n',k}$ at unequal times $n \neq n'$.

In future weeks we will consider spaces of paths that depend on continuous time t rather than discrete time n. The corresponding path spaces, and probability distributions on them, are one of the main subjects of this course.

3 Multivariate normals

Most of the material in this section should be review for most of you. The multivariate Gaussian, or normal, probability distribution is important for so many reasons that it would be dull to list them all here. That activity might help you later as you review for the final exam. The important takeaway is linear algebra as a way to deal with multivariate normals.

3.1 Linear transformations of random variables

sub:lt

Let $X \in \mathbb{R}^d$ be a multivariate random variable. We write $X \sim u(x)$ to indicate that u is the probability density of X. Let A be a square $d \times d$ matrix that describes an invertible linear transformation of random variables $X = AY$, $Y = A^{-1}X$. Let $v(y)$ be the probability density of Y. The relation between u and v is

$$v(y) \;=\; |\det(A)|\; u(Ay) \;. \tag{3}$$

uvd

This is equivalent to

$$u(x) \;=\; \frac{1}{|\det(A)|} x\big(A^{-1}x\big) \;. \tag{4}$$

vud

We will prove it in the form ($\overset{\texttt{uvd}}{3}$) and use it in the form ($\overset{\texttt{vud}}{4}$).

The determinants may be the most complicated things in the formulas ($\overset{\texttt{uvd}}{3}$) and ($\overset{\texttt{vud}}{4}$). They may be the least important. It is common that probability densities are known only *up to a constant.* That is, we know $u(x) = cf(x)$ with a formula for f, but we do not know c. Even if there is a formula for c, the formula may be more helpful without it.

(For example, the Student t-density is

$$u(x) \;=\; c\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} \quad ,$$

with

$$c \;=\; \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \;,$$

in terms of the Euler *Gamma* function $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$. The important features of the t-distribution are easier to see without the formula for c: the fact that it is approximately normal for large n, that it is symmetric and smooth, and that $u(x) \sim x^{-p}$ for large x with exponent $p = n+1$ (power law tails).)

Here is an informal way to understand the transformation rule (uvd 3) and get the determinant prefactor in the right place. Consider a very small region, B_y, in $\mathbb{R}^d$ about the point y. This region could be a small ball or box, say. Call the volume of B_y, informally, $|dy|$. Under the transformation $y \to x = Ay$, say that B_y is transformed to a small region B_x about x. Let $|dx|$ be the volume of B_x. Since $B_y \xrightarrow{A} B_x$ (this means that the transformation A takes B_y to B_x), the ratio of the volumes is given by the determinant:

$$|dx| \;=\; |\det(A)|\;|dy| \;.$$

This formula is exactly true even if $|dy|$ is not small.

But when B_y small, we have the approximate formula

$$\Pr(B_y) \;=\; \int_{B_y} y(y')dy' \;\approx\; v(y)\,|dy| \;. \tag{5}$$

udx

The exact formula $\Pr(B_x) = \Pr(B_y)$ then gives the approximate formula

$$v(y)\,|dy| \;\approx\; u(x)\,|dx| \;=\; u(Ay)\,|\det(A)|\,|dy| \;.$$

In the limit $|dy| \to 0$, the approximations become exact. Cancel the common factor $|dy|$ from both sides and you get the transformation formula (uvd 3).

3.2 Linear algebra, matrix multiplication

sub:la

Very simple facts about matrix multiplication make the mathematician's work much simpler than it would be otherwise. This applies to the *associativity* property of matrix multiplication and the *distributive* property of matrix multiplication and addition. This is part of what makes linear algebra so useful in practical probability.

Suppose A, B, and C are three matrices that are compatible for multiplication. *Associativity* is the formula $(AB)\,C = A\,(BC)$. We can write the product simply as ABC because the order of multiplication does not matter. Associativity holds for products of more factors. For example $(A\,(BC))\,D = (AB)\,(CD)$

gives two of the many ways to calculate the matrix product $ABCD$: you can compute BC, then multiply from the left by A and lastly multiply from the right by D, or you can first calculate AB and CD and then multiply those.

Distributivity is the fact that matrix product is a *linear* function of each factor. Suppose AB is compatible for matrix multiplication, that B_1 and B_2 have the same shape (number of rows and columns) as B, and that u_1 and u_2 are numbers. Then $A(u_1B_1 + u_2B_2) = u_1(AB_1) + u_2(AB_2)$. This works with more than two B matrices, and with matrices on the right and left, such as

$$A\left(\sum_{k=1}^{n} u_k B_k\right) C \;=\; \sum_{k=1}^{n} u_k A B_k C \; .$$

It even works with integrals. If $B(x)$ is a matrix function of $x \in \mathbb{R}^d$ and $u(x)$ is a probability density function, then

$$\int \left(AB(x)C\right)\, u(x)\, dx \;=\; A\left(\int B(x)\, u(x)\, dx\right) C \; .$$

This may be said in a more abstract way. If B is a random matrix and A and C are fixed, not random, then

$$\mathrm{E}[ABC] \;=\; A\,\mathrm{E}[B]\; C \; . \tag{6}$$

eabc

Matrix multiplication is associative and linear even when some of the matrices are row vectors or column vectors. These can be treated as $1 \times d$ and $d \times 1$ matrices respectively.

Of course, matrix multiplication is not commutative: $AB \neq BA$ in general. The matrix transpose reverses the order of matrix multiplication: $\left(AB\right)^t = \left(B^t\right)\left(A^t\right)$. Matrix inverse does the same if A and B are square matrices: $\left(AB\right)^{-1} = \left(B^{-1}\right)\left(A^{-1}\right)$. If A and B are not square, it is possible that AB is invertible even though A and B are not.

We illustrate matrix algebra in probability by finding transformation rules for the mean and covariance of a multivariate random variable. Suppose $Y \in \mathbb{R}^d$ is a d component random variable, and $X = AY$. It is not necessary here for A to be invertible or square, as it was in subsection $\overset{\texttt{sub:lt}}{3.1}$. The mean of Y is the d component vector given either in matrix/vector form as $\mu_Y = \mathrm{E}[Y]$, or in component form as $\mu_{Y,j} = \mathrm{E}[Y_j]$. The expected value of Y is

$$\mu_X \;=\; \mathrm{E}[X] \;=\; \mathrm{E}[AY] \;=\; A\,\mathrm{E}[Y] \;=\; A\,\mu_Y \; .$$

We may take A out of the expectation because of the linearity of matrix multiplication, and the fact that Y may be treated as a $d \times 1$ matrix.

Slightly less trivial is the transformation formula for the covariance matrix. The covariance matrix C_Y is the $d \times d$ symmetric matrix whose entries are

$$C_{Y,jk} \;=\; \mathrm{E}\left[\left(Y_j - \mu_{Y,j}\right)\left(Y_k - \mu_{Y,k}\right)\right] \; .$$

The diagonal entries of C_Y are the variances of the components of Y:

$$C_{Y,jj} \;=\; \mathrm{E}\Big[\left(Y_j - \mu_{Y,j}\right)^2\Big] \;=\; \sigma^2_{Y_j} \;.$$

Now consider the $d \times d$ matrix $B(Y) = (Y - \mu_Y)\,(Y - \mu_Y)^t$. The (j,k) entry of B is just $(Y_j - \mu_{Y,j})\,(Y_k - \mu_{Y,k})$. Therefore the covariance matrix may be expressed as

$$C_Y \;=\; \mathrm{E}\Big[(Y - \mu_Y)\,(Y - \mu_Y)^t\Big] \;. \tag{7}$$

cx

eabc

The linearity formula (6), and associativity, give the transformation law for covariances:

$$\begin{aligned}
C_Y \;&=\; \mathrm{E}\Big[(Y - \mu_Y)\,(Y - \mu_Y)^t\Big] \\
&=\; \mathrm{E}\Big[(AY - A\mu_Y)\,(AY - A\mu_Y)^t\Big] \\
&=\; \mathrm{E}\Big[\{A\,(Y - \mu_Y)\}\,\{A\,(Y - \mu_Y)\}^t\Big] \\
&=\; \mathrm{E}\Big[\Big\{A\,(Y - \mu_Y)\Big\}\,\Big\{(Y - \mu_Y)^t\,A^t\Big\}\Big] \\
&=\; \mathrm{E}\Big[A\big\{(Y - \mu_Y)\,(Y - \mu_Y)^t\big\}A^t\Big] \\
&=\; A\,\mathrm{E}\Big[(Y - \mu_Y)\,(Y - \mu_Y)^t\Big]\;A^t \\
C_X \;&=\; A C_Y A^t \;.
\end{aligned} \tag{8}$$

The second to the third line uses distributivity. The third to the fourth uses the property of matrix transpose. The fourth to the fifth is distributivity again. The fifth to the sixth is linearity.

3.3 Gaussian probability density

This subsection and the next one use the multivariate normal probability density. The aim is not to use the formula but to find ways to avoid using it. We use the general probability density formula to prove the important facts about Gaussians. Working with these general properties is simpler than working with probability density formulas. These properties are

- Linear functions of Gaussians are Gaussian. If X is Gaussian and $Y = AX$, then Y is Gaussian.

- Uncorrelated Gaussians are independent. If X_1 and X_2 are two components of a multivariate normal and if $\mathrm{cov}(X_1, X_2) = 0$ then X_1 and X_2 are independent.

- Conditioned Gaussians are Gaussian. If X_1 and X_2 are two compenents of a multivariate normal, then the distribution of X_1, conditioned on knowing the value $X_2 = x_2$, is Gaussian.

In this subsection and the next, we use the formula for the Gaussian probability density to prove these three properties.

Let H be a $d \times d$ matrix that is *SPD* (symmetric and positive definite). Let $\mu = (\mu_1, \ldots \mu_d)^t \in \mathbb{R}^d$ be a vector. If X has the probability density

$$u(x) \;=\; \frac{\sqrt{\det(H)}}{(2\pi)^{d/2}}\, e^{-(x-\mu)^t H (x-\mu)/2} \;. \qquad (9) \qquad \boxed{\texttt{NH}}$$

then we say that X is a multivariate *Gaussian*, or *normal*, with parameters μ and H. The probability density on the right is denoted by $\mathcal{N}(\mu, H^{-1})$. It will be clear soon why it is convenient to use H^{-1} instead of H. We say that X is *centered* if $\mu = 0$. In that case the density is symmetric, $u(x) = u(-x)$.

It is usually more convenient to write the density formula as

$$u(x) \;=\; c\, e^{-(x-\mu)^t H (x-\mu)/2} \;.$$

The value of the *prefactor*,

$$c \;=\; \frac{\sqrt{\det(H)}}{(2\pi)^{d/2}} \;,$$

often is not important. A probability density is Gaussian if it is the exponential of a quadratic function of x.

We give some examples before explaining ($\overset{\texttt{NH}}{9}$) in general. A *univariate* normal has $d = 1$. In that case we drop the vector and matrix notation because μ and H are just numbers. The simplest univariate normal is the *univariate standard normal*, with $\mu = 0$, and $H = 1$. We often use Z to denote a standard normal, so

$$Z \;\sim\; \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2} \;=\; \mathcal{N}(0,1) \;. \qquad (10) \qquad \boxed{\texttt{sn1}}$$

The *cumulative distribution function*, or *CDF*, of the univariate standard normal is

$$N(x) \;=\; \Pr(\, Z < x \,) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-z^2/2}\, dz \;.$$

There is no explicit formula for $N(x)$ but it is easy to calculate numerically. Most numerical software packages include procedures that compute $N(x)$ accurately.

The general univariate density may be written without matrix/vector notation:

$$u(x) \;=\; \frac{\sqrt{h}}{\sqrt{2\pi}} e^{-(x-\mu)^2 h/2} \;. \qquad (11) \qquad \boxed{\texttt{wn}}$$

Simple calculations (explained below) show that the mean is $\mathrm{E}[X] = \mu$, and the variance is

$$\sigma_X^2 \;=\; \mathrm{var}(X) \;=\; \mathrm{E}\Big[(X-\mu)^2\Big] \;=\; \frac{1}{h} \;.$$

In view of this, the probability density ($\overset{\texttt{wn}}{11}$) is also written

$$u(x) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} \;. \qquad (12) \qquad \boxed{\texttt{uvn}}$$

If X has this density, we write $X \sim \mathcal{N}(\mu, \sigma^2)$. It would have been more accurate to write σ_X^2 and μ_X, but we make a habit of dropping subscripts that are easy to guess from context.

It is useful to express a general univariate normal in terms of a standard univariate normal. If we want $X \sim \mathcal{N}(\mu, \sigma^2)$, we can take $Z \sim (0,1)$ and take

$$X \;=\; \mu \,+\, \sigma Z \;. \tag{13} \qquad \boxed{\texttt{XmsZ}}$$

It is clear that $\mathrm{E}[X] = \mu$. Also,

$$\begin{aligned} \mathrm{var}(X) \;&=\; \mathrm{E}\Big[(X-\mu)^2 \Big] \\ &=\; \sigma^2 \, \mathrm{E}\big[Z^2 \big] \;=\; \sigma^2 \;. \end{aligned}$$

A calculation with probability densities (see below) shows that if ($\overset{\texttt{sn1}}{10}$) is the density of Z and X is given by ($\overset{\texttt{XmsZ}}{13}$), then X has the density ($\overset{\texttt{uvn}}{12}$). This is handy in calculations, such as

$$\Pr[X < a] \;=\; \Pr[\mu + \sigma Z < a] \;=\; \Pr[Z < (a-\mu)\,/\sigma] \;=\; N\!\left(\frac{a-\mu}{\sigma} \right) \;.$$

This says that the probability of $X < a$ depends on how many standard deviations a is away from the mean, which is the argument of N on the right.

We move to a multivariate normal and return to matrix and vector notation. The *standard multivariate normal* has $\mu = 0$ and $H = I_{d\times d}$. In this case, the exponent in ($\overset{\texttt{NH}}{9}$) involves just $z^t H z = z^t z = z_1^2 + \cdots z_d^2$, and $\det(H) = 1$. Therefore the $\mathcal{N}(0, I)$ probability density ($\overset{\texttt{NH}}{9}$) is

$$\begin{aligned} Z \;&\sim\; \frac{1}{(2\pi)^{d/2}} \, e^{-z^t z/2} \qquad\qquad (14) \\ &=\; \left(\frac{1}{\sqrt{2\pi}} \right)^{d} e^{-z_1^2 - \cdots - z_d^2/2} \\ &=\; \left(\frac{1}{\sqrt{2\pi}} e^{-z_1^2/2} \right) \cdot \left(\frac{1}{\sqrt{2\pi}} e^{-z_2^2/2} \right) \cdot \;\cdots\; \cdot \left(\frac{1}{\sqrt{2\pi}} e^{-z_d^2/2} \right) \;. \qquad (15) \end{aligned}$$

The last line writes the probability density of Z as a product of one dimensional $\mathcal{N}(0,1)$ densities for the components $Z_1, \ldots, Z_d$. This implies that the components Z_j are independent univariate standard normals. The elements of the covariance matrix are

$$\left. \begin{array}{lclclcl} C_{Z,jj} & = & \mathrm{var}(Z_j) & = & \mathrm{E}\big[Z_j^2 \big] & = & 1 \\ C_{Z,jk} & = & \mathrm{cov}(Z_j, Z_k) & = & \mathrm{E}\big[Z_j Z_k^2 \big] & = & 0 \quad \text{if } j \neq k \end{array} \right\} \;. \tag{16} \qquad \boxed{\texttt{ZI}}$$

In matrix terms, this is just $C_Z = I$. The covariance matrix of the multivariate standard normal is the identity matrix. In this case at least, uncorrelated Gaussian components Z_j are also independent.

The themes of this section so far are the general transformation law ($\overset{\texttt{vud}}{4}$), the covariance transformation formula ($\overset{\texttt{cycx}}{8}$), and the Gaussian density formula ($\overset{\texttt{NH}}{9}$).

We are ready to combine them to see how multivariate normals transform under linear transformations. Suppose Y has probability density (assuming for now that $\mu = 0$)

$$v(y) \;=\; c\,e^{-y^tHy/2}\;,$$

and $X = AY$, with an invertible A so $Y = A^{-1}X$. We use $(\overset{\texttt{vud}}{4})$, and calculate the exponent

$$y^tHy \;=\; \left(A^{-1}x\right)^t H\left(A^{-1}x\right) \;=\; x^t\left[\left(A^{-1}\right)^t HA^{-1}\right]x \;=\; x^t\widetilde{H}x\;,$$

with $\widetilde{H} = \left(A^{-1}\right)^t HA^{-1}$. (Note that $\left(A^{-1}\right)^t = \left(A^t\right)^{-1}$. We denote these by A^{-t}, and write $\widetilde{H} = A^{-t}HA^{-1}$.) The formula for $\widetilde{H}$ is not important here. What is important is that $u(x) = c\,e^{-x^t\widetilde{H}x/2}$, which is Gaussian. This proves that a linear transformation of a multivariate normal is a multivariate normal, at least if the linear transformation is invertible.

We come to the relationship between H, the SPD matrix in $(\overset{\texttt{NH}}{9})$, and C, the covariance matrix of X. In one dimension the relation is $\sigma^2 = C = h^{-1}$. We now show that for $d \geq 1$ the relationship is

$$C \;=\; H^{-1}\;. \qquad (17) \quad \boxed{\texttt{CH}}$$

The multivariate normal, therefore, is $\mathcal{N}(\mu, H^{-1}) = \mathcal{N}(\mu, C)$. This is consistent with the one variable notation $\mathcal{N}(\mu, \sigma^2)$. The relation $(\overset{\texttt{CH}}{17})$ allows us to rewrite the probability density $(\overset{\texttt{NH}}{9})$ in its more familiar form

$$u(x) \;=\; c\,e^{-(x-\mu)^tC^{-1}(x-\mu)/2}\;. \qquad (18) \quad \boxed{\texttt{N}}$$

The prefactor is (presuming $C = H^{-1}$)

$$c \;=\; \frac{1}{(2\pi)^{d/2}\sqrt{\det(C)}} \;=\; \frac{\sqrt{\det(H)}}{(2\pi)^{d/2}}\;. \qquad (19) \quad \boxed{\texttt{pf}}$$

The proof of $(\overset{\texttt{CH}}{17})$ uses an idea that is important for computation. A natural multivariate version of $(\overset{\texttt{XmsZ}}{13})$ is

$$X \;=\; \mu + AZ\;, \qquad (20) \quad \boxed{\texttt{XmAZ}}$$

where $Z \sim \mathcal{N}(0, I)$. We choose A so that $X \sim \mathcal{N}(\mu, C)$. Then we use the transformation formula to find the density formula for X. The desired $(\overset{\texttt{CH}}{17})$ will fall out. The whole thing is an exercise in linear algebra.

The mean μ property is clear, so we continue to take $\mu = 0$. The covariance transformation formula $(\overset{\texttt{cycx}}{8})$, with $C_Z = I$ in place of C_Y, implies that $C_X = AA^t$. We can create a multivariate normal with covariance C if we can find an A with

$$AA^t \;=\; C\;. \qquad (21) \quad \boxed{\texttt{cho}}$$

You can think of A as the square root of C, just as σ is the square root of σ^2 in the one dimensional version $(\overset{\texttt{XmsZ}}{13})$.

There are different ways to find a suitable A. One is the Cholesky factorization $C = LL^t$, where L is a lower triangular matrix. This is described in any good linear algebra book (Strang, Lax, not Halmos). This is convenient for computation because numerical software packages usually include a routine that computes the Cholesky factorization.

This is the algorithm for creating X. We now find the probability density of X using the transformation formula (4). We write c for the prefactor in any probability density formula. The value of c can be different in different formulas. We saw that $v(z) = ce^{-z^t z/2}$. With $z = A^{-1}x$, we get

$$u(x) \;=\; c\,v(A^{-1}x) \;=\; c\,e^{-\left(A^{-1}x\right)^t A^{-1}x/2}\;.$$

The exponent is, as we just saw, $-x^2Hx/2$ with

$$H \;=\; A^{-t}A^{-1}\;.$$

But if we take the inverse of both sides of (21) we find $C^{-1} = A^{-t}A^{-1}$. This proves (17), as the expressions for C^{-1} and H are the same.

The prefactor works out too. The covariance equation (21) implies that

$$\det(C) \;=\; \det(A)\det(A^t) \;=\; \left[\,\det(A)\,\right]^2\;.$$

Using (14) and (3) together gives

$$c \;=\; \frac{1}{(2\pi)^{d/2}}\,\frac{1}{\det(A)}\;,$$

which is the prefactor formula (19).

Two important properties of the multivariate normal, for your review list, is that they exist for any C and are easy to generate. The covariance square root equation (21) has a solution for any SPD matrix C. If C is a desired covariance matrix, the mapping (20) produces multivariate normal X with covariance C. Standard software packages include random number generators that produce independent univariate standard normals Z_k. If you have C and you want a million vectors independent random vectors X, you first compute the Cholesky factor, L. Then a million times you use the standard normal random number generator to produce a d component standard normal Z and do the matrix calculation $X = LZ$.

This makes the multivariate normal family different from other multivariate families. Sampling a general multivariate random variable can be challenging for d larger than about 5. Practitioners resort to heavy handed and slow methods such as *Markov chain Monte Carlo*. Moreover, there is a modeling question that may be hard to answer for general random variables. Suppose you want univariate random variables X_1, ..., X_d each to have density $f(x)$ and you want them to be correlated. If $f(x)$ is a univariate normal, you can make the X_k components of a multivariate normal with desired variances and correlations. If the X_k are not normal, the *copula* transformation maps the situation to a multivariate normal. *Warning*: the copula transformation has been blamed for the 2007 financial meltdown, seriously.

3.4 Conditional and marginal distributions

sub:cm

When you talk about conditional and marginal distributions, you have to say which variables are fixed or known, and which are variable or unknown. We write the multivariate random variable as (X, Y) with $X \in \mathbb{R}^{d_X}$ and $y \in \mathbb{R}^{d_Y}$. The number of random variables in total is still $d = d_X + d_Y$. We study the conditional distribution of X, conditioned on knowing $Y = y$. We also study the marginal distribution of Y.

The math here is linear algebra with block vectors and matrices. The total random variable is *partitioned* into its X and Y parts as

$$\left(\frac{X}{Y}\right) \;=\; \begin{pmatrix} X_1 \\ \vdots \\ X_{d_X} \\ Y_1 \\ \vdots \\ Y_{d_Y} \end{pmatrix} \;\in\; \mathbb{R}^d$$

The Gaussian joint distribution of X and Y has in its exponent

$$\left(x^t, y^t\right) \left(\begin{array}{c|c} H_{XX} & H_{XY} \\ \hline H_{YX} & H_{YY} \end{array}\right) \left(\frac{x}{y}\right) \;=\; x^t H_{XX} x \;+\; 2x^t H_{XY} y \;+\; y^t H_{YY} y \;.$$

(This relies on $x^t H_{XY} y = y^t H_{YX} x$, which is true because $H_{YX} = H^t_{XY}$, which is true because H is symmetric.) The joint distribution of X and Y is

$$u(x, y) \;=\; c\, e^{-\left(x^t H_{XX} x \,+\, 2x^t H_{XY} y \,+\, y^t H_{YY} y\right)/2} \;.$$

If $u(x, y)$ is any joint distribution, then the conditional density of X conditioned on $Y = y$, is $u(x \,|\, Y = y) = u(x \,|\, y) = c(y) u(x, y)$. Here we think of x as the variable and y as a parameter. The normalization constant here depends on the parameter. For the Gaussian, we have

$$u(x \,|\, Y = y) \;=\; c(y)\, e^{-\left(x^t H_{XX} x \,+\, 2x^t H_{XY} y\right)/2} \;. \tag{22}$$

cX

The factor $e^{-y^t H_{YY} y/2}$ has been absorbed into $c(y)$.

We see from (22) that the conditional distribution of X is the exponential of a quadratic function of x, which is to say, Gaussian. The algebraic trick of *completing the square* identifies the conditional mean. We seek to write the exponent (22) in the form of the exponent of (9)

$$x^t H_{XX} x \;+\; 2x^t H_{XY} y \;=\; \left(x - \mu_X(y)\right)^t H_{XX} \left(x - \mu_X(y)\right) \;+\; m(y) \;.$$

The $m(y)$ will eventually be absorbed into the y dependent prefactor. Some algebra shows that this works provided

$$2x^t H_{XY} y \;=\; -2x^t H_{XX}\, \mu_X(y) \;.$$

This will hold for all x if $H_{XY}y \;=\; -H_{XX}\mu_X(y)$, which gives the formula for the conditional mean:

$$\mu_X(y) \;=\; -H_{XX}^{-1}\, H_{XY}\, y \; . \qquad (23)$$ cm

The conditional mean $\mu_X(y)$ is in some sense the best prediction of the unknown X given the known $Y = y$.

3.5 Generating a multivariate normal, interpreting covariance

sub:ch

If we have M with $MM^t = C$, we can think of M as a kind of square root of C. It is possible to find a real $d \times d$ matrix M as long as C is symmetric and positive definite. We will see two distinct ways to do this that give two different M matrices.

The *Cholesky factorization* is one of these ways. The Cholesky factorization of C is a *lower triangular* matrix L with $LL^t = C$ Lower triangular means that all non-zero entries of L are on or below the digonal:

$$L \;=\; \begin{pmatrix} l_{11} & 0 & & \cdots & 0 \\ l_{21} & l_{22} & 0 & & \vdots \\ \vdots & & \ddots & \ddots & \\ & & & & 0 \\ l_{d1} & \cdots & & & l_{dd} \end{pmatrix} \; .$$

Any good linear algebra book explains the basic facts of Cholesky factorization. These are such an L exists as long as C is SPD. There is a unique lower triangular L with positive diagonal entries: $l_{jj} > 0$. There is a straightforward algorithm that calculates L from C using approximately $d^3/6$ multiplications (and the same number of additions).

If you want to generate $X \sim \mathcal{N}(\mu, C)$, you compute the Cholesky factorization of C. Any good package of linear algebra software can do this, including downloadable software LAPACK for C or C++ or FORTRAN programming, and the build in linear algebra facilities in Python, R, and Matlab. To make an X, you need d independent standard normals $Z_1, \ldots, Z_d$. Most packages that generate pseudo-random numbers have a procedure to generate such standard normals. This includes Python, R, and Matlab. To do it in C, C++, FORTRAN, you can use a uniform pseudo-random number generator and then use the *Box Muller* formula to get Gaussians. You assemble the Z_j into a vector $Z = (Z_1, \ldots, Z_d$ and take $X = LZ + \mu$.

Consider as an example the two dimensional case with $\mu = 0$. Here, we want X_1 and X_2 that are jointly normal. It is common to specify $\text{var}(X_1) = \sigma_1^2$, $\text{var}(X_2) = \sigma_1^2$, and the *correlation coefficient*

$$\rho_{12} \;=\; \text{corr}(X_1, X_2) \;=\; \frac{\text{cov}(X_1, X_2)}{\sigma_1 \sigma_2} \;=\; \frac{\text{E}(X_1 X_2)}{\sigma_1 \sigma_2} \; .$$

In this case, the Cholesky factor is

$$L \;=\; \begin{pmatrix} \sigma_1 & 0 \\ \rho_{12}\sigma_2 & \sqrt{1-\rho_{12}^2}\sigma_2 \end{pmatrix} \;. \tag{24}$$ $\boxed{\texttt{L2}}$

The general formula $X = LZ$ becomes

$$\begin{aligned} X_1 &= \sigma_1 Z_1 && (25)\\ X_2 &= \rho_{12}\sigma_2 Z_1 + \sqrt{1-\rho_{12}^2}\,\sigma_2 Z_2 \;. && (26) \end{aligned}$$

It is easy to calculate $\mathrm{E}\left[X_1^2\right] = \sigma_1^2$, which is the desired value. Similarly, because Z_1 and Z_2 are independent, we have

$$\mathrm{var}(X_2) \;=\; \mathrm{E}\big[X_2^2\big] \;=\; \rho_{12}^2\sigma_2^2 \;+\; \left(1-\rho_{12}^2\right)\sigma_2^2 \;=\; \sigma_2^2 \;,$$

which is the desired answer, too. The correlation coefficient is also correct:

$$\mathrm{corr}(X_1,X_2) \;=\; \frac{\mathrm{E}\big[X_2^2\big]}{\sigma_1\sigma_2} \;=\; \frac{\mathrm{E}\left[\sigma_1 Z_1 \rho_{12}\sigma_2 Z_1\right]}{\sigma_1\sigma_2} \;=\; \rho_{12}\,\mathrm{E}\big[Z_1^2\big] \;=\; \rho_{12} \;.$$

You can, and should, verify by matrix multiplication that

$$LL^t \;=\; \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{12} \\ \sigma_1\sigma_2\rho_{12} & \sigma_2^2 \end{pmatrix} \;,$$

which is the desired covariance matrix of $(X_1,X_2)^t$.

We could have turned the formulas ($\overset{\texttt{x1}}{25}$) and ($\overset{\texttt{x2}}{26}$) around as

$$\begin{aligned} X_1 &= \sqrt{1-\rho_{12}^2}\,\sigma_1 Z_1 \;+\; \rho_{12}\sigma_1 Z_2 \;+ \\ X_2 &= \sigma_2 Z_2 \;. \end{aligned}$$

In this version, it looks like X_2 is primary and X_1 gets some of its value from X_2. In ($\overset{\texttt{x1}}{25}$) and ($\overset{\texttt{x2}}{26}$), it looks like X_1 is primary and X_2 gets some of its value from X_1. These two models are equally "valid" in the sense that they product the same observed (X_1,X_2) distribution. It is a good idea to keep this in mind when interpreting regression studies involving X_1 and X_2.

4 Linear Gaussian recurrences

$\boxed{\texttt{sec:lgr}}$

Linear Gaussian linear recurrence relations ($\overset{\texttt{lrr}}{1}$) illustrate the ideas in the previous section. We now know that if V_n is a multivariate normal with mean zero, there is a matrix B so that $V_n = BZ_n$, where $Z_n \sim \mathcal{N}(0,I)$, is a standard multivariate normal. Therefore, we rewrite ($\overset{\texttt{lrr}}{1}$) as

$$X_{n+1} \;=\; AX_n \;+\; BZ_n \;. \tag{27}$$ $\boxed{\texttt{lrz}}$

Since the X_n are Gaussian, we need only describe their means and covariances. This section shows that the means and covariances satisfy recurrence relations derived from ($\overset{\texttt{lrr}}{1}$). The next section explores the distributions of paths. This determines, for example, the joint distribution of X_n and X_m for $n \neq m$. These and more general path spaces and path distributions are important throughout the course.

sub:pdd

4.1 Probability distribution dynamics

As long as Z_n is independent of X_n, we can calculate recurrence relations for $\mu_n = \mathrm{E}[X_n]$ and $C_n = \mathrm{cov}[X_n]$. For the mean, we have (you may want to glance back to subsection sub:la 3.2)

$$\begin{aligned} \mu_{n+1} &= \mathrm{E}[AX_n \,+\, BZ_n] \\ &= A\,\mathrm{E}[X_n] \,+\, B\,\mathrm{E}[Z_n] \\ \mu_{n+1} &= A\mu_n \;. \end{aligned} \tag{28}$$

This says that the recurrence relation for the means is the same as the recurrence relation (lrz 27) for the random states if you "turn off the noise" (set Z_n to zero).

For the covariance, it is convenient to combine (lrz 27) and (mur 28) into

$$X_{n+1} \,-\, \mu_{n+1} \;=\; A\,(X_n \,-\, \mu_n) \,+\, BZ_n \;.$$

The covariance calculation starts with

$$\begin{aligned} C_{n+1} &= \mathrm{E}\Big[(X_{n+1} - \mu_{n+1})\,(X_{n+1} - \mu_{n+1})^t\Big] \\ &= \mathrm{E}\Big[(A\,(X_n - \mu_n) + BZ_n)\,(A\,(X_n - \mu_n) + BZ_n)^t\Big] \end{aligned}$$

We expand the last into a sum of four terms. Two of these are zero, one being

$$\mathrm{E}\Big[\Big(A\,(X_n - \mu_n)\,(BZ_n)^t\Big)\Big] \;=\; 0\;,$$

because Z_n has mean zero and is independent of X_n. We keep the non-zero terms:

$$\begin{aligned} C_{n+1} &= \mathrm{E}\Big[(A\,(X_n - \mu_n))\,(A\,(X_n - \mu_n))^t\Big] \,+\, \mathrm{E}\Big[(BZ_n)\,(BZ_n)^t\Big] \\ &= \mathrm{E}\Big[A\,\Big\{(X_n - \mu_n)\,(X_n - \mu_n)^t\Big\}\,A^t\Big] \,+\, \mathrm{E}\big[B\,\big(Z_n Z_n^t\big)\,B^t\big] \\ &= A\,\mathrm{E}\Big[(X_n - \mu_n)\,(X_n - \mu_n)^t\Big]\,A^t \,+\, B\,\mathrm{E}\big[Z_n Z_n^t B^t\big]\,B^t \\ C_{n+1} &= AC_nA^t \,+\, BB^t\;. \end{aligned} \tag{29}$$

The recurrence relations (mur 28) and (cr 29) determine the distribution of X_{n+1} in terms of the distribution of X_n. As such, they are the first example in this class of a *forward equation.*

We will see in subsection sub:ho 4.2 that there are natural examples where the dimension of the noise vector Z_n is less than d, and the noise matrix B is not square. When that happens, we let m denote the number of components of Z_n, which is the number of *sources of noise.* The noise matrix B is $d \times m$; it has d rows and m columns. The case $m > d$ is not important for applications. The matrices in (cr 29) all are $d \times d$, including BB^t. If you wonder whether it might be B^tB instead, note that B^tB is $m \times m$, which might be the wrong size.

4.2 Higher order recurrence relations, the Markov property

sub:ho

It is common to consider recurrence relations with more than one lag. For example, a k lag relation might take the form

$$X_{n+1} \;=\; A_0 X_n \,+\, A_1 X_{n-1} \,+\, \cdots \,+\, A_{k-1} X_{n-k+1} \,+\, BZ_n \;. \qquad (30)$$ lk

From the point of view of X_{n+1}, the k lagged states are X_n (one lag), up to X_{n-k+1} (k lags). It is natural to consider models with multiple lags if X_n represent observable aspects of a large and largely unobservable system. For example, the components of X_n could be public financial data at time n. There is much unavailable private financial data. The lagged values X_{n-j} might give more insight into the complete state at time n than just X_n.

We do not need a new theory of lag k systems. *State space expansion* puts a multi-lag system of the form (30) into the form of a two term recurrence relation (27). This formulation uses expanded vectors

$$\widetilde{X}_n \;=\; \begin{pmatrix} X_n \\ X_{n-1} \\ \vdots \\ X_{n-k+1} \end{pmatrix} \;.$$

If the original states X_n have d components, then the expanded states $\widetilde{X}_n$ have kd components. The noise vector Z_n does not need expanding because noise vectors have no memory. All the memory in the system is contained in $\widetilde{X}_n$. The recurrence relations in the expanded state formulation are

$$\widetilde{X}_{n+1} \;=\; \widetilde{A}\widetilde{X}_n \,+\, \widetilde{B} Z_n \;.$$

In more detail, this is

$$\begin{pmatrix} X_{n+1} \\ X_n \\ \vdots \\ X_{n-k+2} \end{pmatrix} \;=\; \begin{pmatrix} A_0 & A_1 & \cdots & & A_{k+1} \\ I & 0 & \cdots & & 0 \\ 0 & I & \cdots & & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & & I & 0 \end{pmatrix} \begin{pmatrix} X_n \\ X_{n-1} \\ \vdots \\ X_{n-k+1} \end{pmatrix} \;+\; \begin{pmatrix} B \\ 0 \\ \vdots \\ 0 \end{pmatrix} Z_n \;. \qquad (31)$$ cmrr

The matrix $\widetilde{A}$ is the *companion matrix* of the recurrence relation (30).

We will see in subsection 4.3 that the stability of a recurrence relation (27) is determined by the eigenvalues of A. For the case $d = 1$, you might know that the stability of the recurrence relation (30) is determined by the roots of the *characteristic polynomial* $p(z) = z^k - A_0 z^{k-1} - \cdots - A_{k-1}$. These statements are consistent because the roots of the characteristic polynomial are the eigenvalues of the companion matrix.

If X_n satisfies a k lag recurrence (30), then the covariance matrix, $\widetilde{C}_n = \mathrm{cov}(\widetilde{X}_n)$, satisfies $\widetilde{C}_{n+1} = \widetilde{A}\widetilde{C}_n\widetilde{A}^t + \widetilde{B}\widetilde{B}^t$. The simplest way to find the $d \times d$

covariance matrix C_n, is to find the $kd \times kd$ covariance matrix $\widetilde{C}_n$ and look at the top left $d \times d$ block.

The *Markov property* will be important throughout the course. If the X_n satisfy the one lag recurrence relation ($\overset{\texttt{lrz}}{27}$), then they have the Markov property. In this case the X_n form a *Markov chain.* If they satisfy the k lag recurrence relation with $k > 1$ (in a non-trivial way) then the stochastic process X_n does not have the Markov property. The informal definition is as follows. The process has the *Markov property* if X_n is all the information about the past that is relevant for predicting the future. Said more formally, the distribution of X_{n+1} conditional on $X_n, \ldots, X_0$ is the same as the distribution of X_{n+1} conditional on X_n alone.

If a random process does not have the Markov property, you can blame that on the state space being too small, so that X_n does not have as much information about the state of the system as it should. In many such cases, a version of state space expansion can create a more complete collection of information at time n.

Genuine state space expansion, with $k > 1$, always gives a noise matrix $\widetilde{B}$ with fewer sources of noise than state variables. The number of state variables is kd and the number of noise variables is $m \leq d$.

4.3 Large time behavior and stability

sub:ss

Large time behavior is the behavior of X_n as $n \to \infty$. The stochastic process ($\overset{\texttt{lrz}}{27}$) is *stable* if it settles into a stochastic steady state for large n. The states X_n can not have a limit, because of the constant influence of random noise. But the probability distributions, $u_n(x)$, with $X_n \sim u_n(x)$, can have limits. The limit $u(x) = \lim_{n\to\infty} u_n(x)$ is a *statistical steady state.* The finite time distributions u_n are Gaussian: $u_n = \mathcal{N}(\mu_n, C_n)$, with μ_n and C_n satisfying the recurrences ($\overset{\texttt{mur}}{28}$) and ($\overset{\texttt{cr}}{29}$). The limiting distribution depends on the following limits:

$$\mu = \lim_{n\to\infty} \mu_n \tag{32}$$

$$C = \lim_{n\to\infty} C_n \tag{33}$$

If these limits exist, then[3] $u = \mathcal{N}(\mu, C)$.

In the following discussion we first ignore several subtleties in linear algebra for the sake of simplicity. Conclusions are correct as initially stated if $m = d$, B is non-singular, and there are no Jordan blocks in the eigenvalue decomposition of A. We will then re-examine the reasoning to figure out what can happen in exceptional degenerate cases.

The limit ($\overset{\texttt{mul}}{32}$) depends on the eigenvalues of A. Denote the eigenvalues by λ_j and the corresponding right eigenvectors by r_j, so that $Ar_j = \lambda_j r_j$ for $j = 1, \ldots, d$. The eigenvalues and eigenvectors do not have to be real even when A is real. The eigenvectors form a basis of $\mathbb{C}^d$, so the means μ_n have

[3] Some readers will worry that this statement is not proven with mathematical rigor. It can be, but we are avoiding that kind of technical discussion.

unique representations $\mu_n = \sum_{j=1}^d m_{n,j} r_j$. The dynamics ($\overset{\texttt{mur}}{28}$) implies that $m_{n+1,j} = \lambda_j m_{n,j}$. This implies that

$$m_{n,j} \;=\; \lambda_j^n m_{0,j} \; . \tag{34} \quad \boxed{\texttt{mp}}$$

The matrix A is *strongly stable* if $|\lambda_j| < 1$ for $j = 1, \ldots, d$. In this case $m_{n,j} \to 0$ as $n \to \infty$ for each j. In fact, the convergence is *exponential*. We see that if A is strongly stable, then $\mu_n \to 0$ as $n \to \infty$ independent of the initial mean μ_0. The opposite case is that $|\lambda_j| > 1$ for some j. Such an A is *strongly unstable*. It usually happens that $|\mu_n| \to \infty$ as $n \to \infty$ for a strongly unstable A. The limiting distribution u does not exist for strongly unstable A. The borderline case is $|\lambda_j| \leq 1$, for all j and there is at least one j with $|\lambda_j| \leq 1$. This may be called either *weakly stable* or *weakly unstable*.

If A is strongly stable, then the limit ($\overset{\texttt{c1}}{33}$) exists. We do not expect $C_n \to 0$ because the uncertainty in X_n is continually replenished by noise. We start with a direct but possibly unsatisfying proof. A second and more complicated proof follows. The first proof just uses the fact that if A is strongly stable, then

$$\|A^n\| \;\leq\; c\, a^n \; , \tag{35} \quad \boxed{\texttt{ab}}$$

for some constant c and positive $a < 1$. The value of c depends on the matrix norm and is not important for the proof.

We prove that the limit ($\overset{\texttt{c1}}{33}$) exists by writing C as a convergent infinite sum. To simplify notation, write R for BB^t. Suppose C_0 is given, then ($\overset{\texttt{cr}}{29}$) gives $C_1 = AC_0A^t + R$. Using ($\overset{\texttt{cr}}{29}$) again gives

$$\begin{aligned} C_2 \;&=\; AC_1A^t + R \\ &=\; A\left(AC_0A^t + R\right)A^t + R \\ &=\; A^2C_0\left(A^t\right)^2 + ARA^t + R \\ &=\; A^2C_0\left(A^2\right)^t + ARA^t + R \end{aligned}$$

We can continue in this way to see (by induction) that

$$C_n = A^nC_0\left(A^n\right)^t + A^{n-1}R\left(A^{n-1}\right)^t + \cdots + R \; .$$

This is written more succinctly as

$$C_n \;=\; A^nC_0\left(A^n\right)^t + \sum_{k=0}^{n-1} A^kR\left(A^k\right)^t \; . \tag{36} \quad \boxed{\texttt{gsf}}$$

The limit of the C_n exists because the first term on the right goes to zero as $n \to \infty$ and the second term converges to the infinite sum

$$C \;=\; \sum_{k=0}^{\infty} A^kR\left(A^k\right)^t \; . \tag{37} \quad \boxed{\texttt{gsi}}$$

For the first term, note that (35) and properties of matrix norms imply that[4]

$$\left\| A^n C_0 \left(A^n\right)^t \right\| \;\leq\; (ca^n)\, \|C_0\|\, (ca^n) \;=\; ca^{2n} \|C_0\| \; .$$

We write c instead of c^2 at the end because c is a generic constant whose value does not matter. The right side goes to zero as $n \to \infty$ because $a < 1$. For the second term, recall that an infinite sum is the limit of its partial sums if the infinite sum converges absolutely. Absolute convergence is the convergence of the sum of the absolute values, or the norms in case of vectors and matrices. Here the sum of norms is:

$$\sum_{k=0}^{\infty} \left\| A^k R \left(A^k\right)^t \right\| \; .$$

Properties of norms bound this by a geometric series:

$$\left\| A^k R \left(A^k\right)^t \right\| \;\leq\; c\, a^{2k}\, \|R\| \; .$$

You can find C without summing the infinite series (37). Since the limit (33) exists, you can take the limit on both sides of (29), which gives

$$C \,-\, ACA^t \;=\; BB^t \; . \tag{38}$$

le

Subsection 4.4 explains that this is a system of linear equations for the entries of C. The system is solvable and the solution is positive definite if A is strongly stable. As a warning, (38) is solvable in most cases even when A is strongly unstable. But in those cases the C you get is not positive definite and therefore is not the covariance matrix of anything. The dynamical equation (29) and the steady state equation (38) are examples of *Liapounov equations.*

Here are the conclusions: if A is strongly stable then u_n, the distribution of X_n has $u_n \to u$ as $n \to \infty$, with a Gaussian limit $u = \mathcal{N}(0, C)$, and C is given by (37), or by solving (38). If A is not strongly stable, then it is unlikely that the u_n have a limit as $n \to \infty$. It is not altogether impossible in degenerate situations described below. If A is strongly unstable, then it is most likely that $\|\mu_n\| \to \infty$ as $n \to \infty$. If A is weakly unstable, then probably $\|C_n\| \to \infty$ as $n \to \infty$ because the sum (37) diverges.

sub:ev

4.4 Linear algebra and the limiting covariance

This subsection is a little esoteric. It is (to the author) interesting mathematics that is not strictly necessary to understand the material for this week. Here we find eigenvalues and *eigen-matrices* for the recurrence relation (29). These are related to the eigenvalues and eigenvectors of A.

[4] Part of this expression is similar to the design on Courant Institute tee shirts.

The covariance recurrence relation (29)has the same stability/instability dichotomy. We explain this by reformulating it as more standard linear algebra. Consider first the part that does not involve B, which is

$$C_{n+1} \;=\; AC_nA^t \;. \tag{39}$$

Le

Here, the entries of C_{n+1} are linear functions of the entries of C_n. We describe this more explicitly by collecting all the distinct entries of C_n into a vector $\vec{c}_n$. There are $D = (d+1)d/2$ entries in $\vec{c}_n$ because the elements of C_n below the diagonal are equal to the entries above. For example, for $d = 3$ there are $D = 6$ distinct entries in C_n, which are $C_{n,11}$, $C_{n,12}$, $C_{n,13}$, $C_{n,22}$, $C_{n,23}$, and $C_{n,33}$, which makes $\vec{c}_n = (C_{n,11}, C_{n,12}, C_{n,13}, C_{n,22}, C_{n,23}, C_{n,33})^t \in \mathbb{R}^D (= \mathbb{R}^6)$. There is a $D \times D$ matrix, L so that $\vec{c}_{n+1} = L\vec{c}_n$. In the case $d = 2$ and $A = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$, the C_n recurrence relation, or dynamical Liapounov equation without BB^t, (29) is

$$\begin{pmatrix} C_{n+1,11} & C_{n+1,12} \\ C_{n+1,12} & C_{n+1,22} \end{pmatrix} \;=\; \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} C_{n+1,11} & C_{n+1,12} \\ C_{n+1,12} & C_{n+1,22} \end{pmatrix} \begin{pmatrix} \alpha & \gamma \\ \beta & \delta \end{pmatrix} \;.$$

This is equivalent to $D = 3$ and

$$\begin{pmatrix} C_{n+1,11} \\ C_{n+1,12} \\ C_{n+1,22} \end{pmatrix} \;=\; \begin{pmatrix} \alpha^2 & 2\alpha\beta & \beta^2 \\ \alpha\gamma & \beta\gamma + \alpha\delta & \beta\delta \\ \gamma^2 & 2\gamma\delta & \delta^2 \end{pmatrix} \begin{pmatrix} C_{n,11} \\ C_{n,12} \\ C_{n,22} \end{pmatrix} \;.$$

And that identifies L as

$$L \;=\; \begin{pmatrix} \alpha^2 & 2\alpha\beta & \beta^2 \\ \alpha\gamma & \beta\gamma + \alpha\delta & \beta\delta \\ \gamma^2 & 2\gamma\delta & \delta^2 \end{pmatrix} \;.$$

This formulation is not so useful for practical calculations. Its only purpose is to show that (39) is related to a $D \times D$ matrix L.

The limiting behavior of C_n depends on the eigenvalues of L. It turns out that these are determined by the eigenvalues of A in a simple way. For each pair (j,k) there is an eigenvalue of L, which we call μ_{jk}, that is equal to $\lambda_j\lambda_k$. To understand this, note that an eigenvector, $\vec{s}$, of L, with $L\vec{s} = \mu\vec{s}$, corresponds to a symmetric $d \times d$ eigen-matrix, S, with

$$ASA^t \;=\; \mu S \;.$$

It happens that $S_{jk} = r_j r_k^t + r_k r_j^t$ is the eigen-matrix corresponding to eigenvalue $\mu_{jk} = \lambda_i\lambda_j$. (To be clear, S_{jk} is a $d \times d$ matrix, not the (j,ik) entry of a matrix

S.) For one thing, it is symmetric ($S_{jk}^t = S_{jk}$). For another thing:

$$\begin{aligned}
AS_{jk}A^t &= A\left(r_jr_k^t + r_kr_j^t\right)A^t \\
&= A\left(r_jr_k^t\right)A^t \,+\, A\left(r_kr_j^t\right)A^t \\
&= \left(Ar_j\right)\left(Ar_k\right)^t \,+\, \left(Ar_k\right)\left(Ar_j\right)^t \\
&= \left(\lambda_jr_j\right)\left(\lambda_kr_k\right)^t \,+\, \left(\lambda_kr_k\right)\left(\lambda_jr_j\right)^t \\
&= \lambda_j\lambda_j\left(r_jr_k^t + r_kr_j^t\right) \\
&= \mu_{jk}S_{jk}\ .
\end{aligned}$$

A counting argument shows that all the eigenvalues and eigen-matrices of L take the form of S_{jk} for some $j \geq k$. The number of such pairs is the same D, which is the number of independent entries in a general symmetric matrix. We do not count S_{jk} with $j < k$ because $S_{jk} = S_{kj}$ with $k > j$.

Now suppose A is strongly stable. Then the Liapounov dynamical equation (cr 29) is equivalent to

$$\vec{c}_{n+1} \;=\; L\vec{c}_n \,+\, \vec{r}\ .$$

Since all the eigenvalues of L are less than one in magnitude, a little reasoning with linear algebra shows that $\vec{c}_n \to \vec{c}$ as $n \to \infty$, and that $\vec{c} - L\vec{c} = (I - L)\,\vec{c} = \vec{r}$. The matrix $I - L$ is invertible because L has no eigenvalues equal to 1. This is a different proof that the steady state Liapounov equation (le 38) has a unique solution. It is likely that L has no eigenvalue equal to 1 even if A is not strongly stable. In this case (le 38) has a solution, which is a symmetric matrix C. But there is no guarantee that this C is positive definite, so it does not represent a covariance matrix.

sub:d

4.5 Degenerate cases

The simple conclusions of subsections sub:ss 4.3 and sub:ev 4.4 do not hold in every case. The reasoning there assumed things about the matrices A and B that you might think are true in almost every interesting case. But it is important to understand how things might more complicated in borderline and degenerate cases. For one thing, many important special cases are such borderline cases. Many more systems have behavior that is strongly "influenced" by near degeneracy. A process that is weakly but not strongly unstable is simple Gaussian random walk, which is a model of Brownian motion. A covariance that is nearly singular is the covariance matrix of asset returns, of the S&P 500 stocks. This is a matrix of rank 500 that is pretty well approximated for many purposes by a matrix of rank 10.

4.5.1 Rank of B

The matrix B need not be square or have rank d.

5 Paths and path space

sec:p

There are questions about the process ($\overset{\texttt{lrr}}{1}$) that depend X_n for one n. For example, what is $\Pr\left(\|X_n\| \leq 1 \text{ for } 1 \leq n \leq 10\right)$? The probability distribution on path space answers such questions. For linear Gaussian processes, the distribution in path space is Gaussian. This is not surprising. This subsection goes through the elementary mechanics of Gaussian path space. We also describe more general path space terminology that carries over to other other kinds of Markov processes.

Two relevant probability spaces are the *state space* and the *path space*. We let $\mathcal{S}$ denote the state space. This is the set of all possible values of the state at time n. This week, the state is a d component vector and $\mathcal{S} = \mathbb{R}^d$. The path space is called Ω. This week, Ω is sequences of states with a given starting and ending time. That is, $X_{[n_1:n_2]} \in \Omega$ is a sequence $(X_{n_1}, X_{n_1+1}, \ldots, X_{n_2})$. There are $n_2 - n_1 + 1$ states in the sequence, so $\Omega = \mathbb{R}^{(n_2-n_1+1)d}$. Even if the state space is not $\mathbb{R}^d$, still a path is a sequence of states. We express this by writing $\Omega = \mathcal{S}^{n_2-n_1+1}$. The path space Ω depends on n_1 and n_2 (only the difference, really), but we leave that out of the notation because it is usually clear from the discussion.

6 Exercises

sec:ex

ex:cg

1. This exercise works through conditional distributions of multivariate normals in a sequence of steps. The themes (for the list of facts about Gaussians) are the role of linear algebra and the relation to linear regression. Suppose X and Y have d_X and d_Y components respectively. Let $u(x,y)$ be the joint density. Then the conditional distribution of Y conditional on $X = x$ is $u(\,y \,|\, X = x) = c(x)u(x,y)$. This says that the conditional distribution is the same, up to a normalization constant) as the joint distribution once you fix the variable whose value is known (x in this case). The normalization constant is determined by the requirement that the conditional distribution has total probability equal to 1:

   $$c(x) \;=\; \frac{1}{\int u(x,y)\,dy}\;.$$

   For Gaussian random variables, finding $c(x)$ usually both easy and unnecessary.

   (a) This part works out the simplest case. Take $d = 2$, and $X = (X_1, X_2)^t$. Suppose $X \sim \mathcal{N}(0, H^{-1})$. Fix the value of $X_1 = x_1$ and calculate the distribution of the one dimensional random variable X_2. If H is

   $$H \;=\; \begin{pmatrix} h_{11} & h_{12} \\ h_{12} & h_{22} \end{pmatrix}\;,$$

then the joint density is

$$u(x_1,x_2) \;=\; c\,\exp\!\left[-\left(h_{11}x_1^2 \,+\, 2h_{12}x_1x_2 \,+\, h_{22}x_2^2\right)/2\right] \;.$$

The conditional density looks almost the same:

$$u(\,x_2 \,|\, x_1) \;=\; c(x_1)\,\exp\!\left[-\left(2h_{12}x_1x_2 \,+\, h_{22}x_2^2\right)/2\right] \;.$$

Why is it allowed to leave the term $h_{11}x_1^2$ out of the exponent? Complete the square to write this in the form

$$u(\,x_2 \,|\, x_1) \;=\; c(x_1)\,\exp\!\left[-\left(x_2 - \mu(x_1)\right)^2/(2\sigma_2^2)\right] \;.$$

Find formulas for the conditional mean, $\mu(x_1)$, and the conditional variance, σ_2^2.