Stochastic Calculus, Courant Institute, Fall 2015
http://www.math.nyu.edu/faculty/goodman/teaching/StochCalc2015/index.html

Section 1, Discrete time Gaussian processes

Jonathan Goodman, September, 2015

1 Review and notation

Much of the material here should be familiar to many readers, though likely not in exactly the way it is presented here. Students in Stochastic Calculus form a very heterogeneous group. You have a variety of educational backgrounds, time away from school, and goals. A bit of review will establish a common system of terminology for the class and it will fill in gaps in your background.

Gaussian random variables are central to stochastic calculus, for many reasons. The most important stochastic process for this class is Brownian motion, which is Gaussian. Brownian motion itself is an important model for many physical and financial processes. A much larger class of processes, *diffusions processes* behave like Brownian motion and represented using Brownian motion. Discrete time Gaussian processes, discussed in this section, are another important class of models. The many special properties of Gaussian random variables allow Gaussian process models to be "solved": many of their properties can be calculated explicitly. Gaussian processes play a role similar to the role played by linear systems in the study of dynamical systems. Most processes are not linear/Gaussian, but linear/Gaussian analysis is one of our most useful analytical tools.

1.1 Probability densities and linear transformations

When we study stochastic processes, we usually have many random variables that are related to each other in some way. For example, X_1, X_2, ..., could be the values of some random quantity observed at different times. We may call the individual values X_k the *components* of the *multivariate* random variable $X = (X_1, \ldots, X_d)$. This class focuses mostly (but not entirely, see Section 2), on random variables with continuous joint probability densities. A multivariate random variable, X, is *Gaussian* if its PDF (*Probability Density Function*) has a specific form, see (9). This is not only a condition on the individual components X_k, but on the joint distribution. We sometimes say say that the components are *jointly Gaussian*, or *jointly normal* to emphasize not only that they are normal, but also they are the components of a multivariate normal.

It is possible that the components of X are Gaussian but X is not Gaussian. If the components X_k are jointly Gaussian, then a *linear combination* of them is also Gaussian. If $m_1, \ldots, m_d$ are some factors, a linear combination is $Y = m_1X_1 + \cdots + m_dX_d$. The numbers m_k may also be called *multipliers* or *weights* depending on the context. It is a theorem that the components X_k are jointly

Gaussian, then any linear combination, Y is also Gaussian. We will see this soon. It is also true, but harder to show, that if every linear combination is Gaussian, then the X_k are jointly Gaussian. Linear combinations are common in multivariate probability and stochastic calculus. Some of them are suggested by application, such as the sum of the components $Y = X_1 + \cdots + X_d$. This is a linear combination with equal weights. Others are carefully constructed to have useful mathematical properties, such as *principal components* discussed below.

Just as we often consider more than one component at a time, we often consider more than one linear combination at a time. We can consider a collection of linear combinations

$$Y_j = \sum_{k=1}^{d} m_{jk} X_k \ .$$

The combinations Y_j form the components of a multivariate random variable, $Y = (Y_1, \ldots, Y_n)$. Each of the components Y_j has its own set of weights m_{jk}. One common example is the *partial sums*

$$\begin{aligned} Y_1 &= X_1 \\ Y_2 &= X_1 + X_2 \\ &\vdots \\ Y_d &= X_1 + \cdots + X_d \ . \end{aligned}$$

The reader probably is familiar with the fact that matrices and linear transformations are the "right way" to talk about collections of linear transformations.

Suppose X_1, ..., X_d is a collection of random variables with some joint density. We write

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

for the vector in $\mathbb{R}^d$ with these random variables as components. The corresponding *probability density function*, or *PDF*, is $u(x)$, with $x \in \mathbb{R}^d$.

Suppose M is a non-singular $d \times d$ matrix and $Y = MX$ has PDF $v(y)$. The relation between the two probability densities can be written in two equivalent ways

$$v(y) = \left|\det(M)\right|^{-1} u(M^{-1}y) \tag{1}$$

$$v(Mx) = \left|\det(M)\right| u(x) \ . \tag{2}$$

This formula follows from the general form for changes of variable in multidimensional integration. To explain it informally, suppose A is small set near x_0 whose volume is $|dx|$. Then

$$\Pr(X \in A) = u(x_0) \left|dx\right| \ .$$

This formula is not exactly true for finite sized A, but becomes more and more accurate approximation as A becomes smaller around x_0. Suppose $B = MA$, which means B is the *image* of A under M. This means that $y \in B$ if there is an $x \in A$ with $Mx = y$. Because $x \to Mx$ is a linear transformation, even if A is not small we have

$$\mathrm{vol}(B) = |\det(M)| \, \mathrm{vol}(A) \ .$$

If $Y = MX$, then $\Pr(Y \in B) = \Pr(X \in A)$, so, if $y_0 = Mx_0$, then

$$\begin{aligned} v(y_0)\,|dy| &= u(x_0)\,|dx| \\ v(Mx_0)\,|\det(M)|\,|dx| &= u(x_0)\,|dx| \ . \end{aligned}$$

This is the transformation formula (1).

Determinants can be hard to evaluate in general, but some are easy. If M is a diagonal matrix,

$$M = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & \lambda_d \end{pmatrix} \ ,$$

then

$$\det(M) = \prod_{j=1}^{d} \lambda_j \ .$$

If M is a lower triangular matrix

$$M = \begin{pmatrix} m_{11} & 0 & \cdots & 0 \\ m_{21} & m_{22} & 0 & \vdots \\ \vdots & & \ddots & \\ m_{d1} & m_{d2} & \cdots & m_{dd} \end{pmatrix} \ ,$$

then there is just one term in the determinant formula, which is the product of the diagonals:

$$\det(M) = \prod_{j=1}^{d} m_{jj} \ .$$

For example, the partial sums above are described by the matrix

$$M = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \vdots \\ \vdots & & \ddots & \\ 1 & 1 & \cdots & 1 \end{pmatrix}$$

The determinant of this matrix is 1. The determinant of a matrix is equal to the determinant of its transpose

$$\det(M) = \det(M^t) \ .$$

The determinant of a product is the product of the determinants

$$\det(MN) = \det(M)\det(N) \ .$$

This applies also if there are more than two factors. For example, if A is a symmetric matrix that is diagonalized by the orthogonal matrix V,

$$A = V\Lambda V^t \ \ , \quad VV^t = I \ ,$$

then $\det(V) = \pm 1$, because $\det(V)$ is a real number with

$$[\det(V)]^2 = \det(V)\det(V^t) = \det(VV^t) = \det(I) = 1 \ .$$

Therefore

$$\det(A) = \det(V\Lambda V^t) = \det(V)\det(\Lambda)\det(V^t) = \det(\Lambda) = \prod_{j=1}^{d} \lambda_j \ .$$

If A is symmetric and positive definite, and has the Choleski factorization $A = LL^t$ (more on Choleski below), with L being lower triangular, then

$$\det(A) = [\det(L)]^2 = \prod_{j=1}^{d} l_{jj}^2 \ .$$

1.2 Matrices and linear algebra

Simple facts about matrix multiplication make the mathematician's work much simpler than it would be otherwise.[1] Among these facts are the *associativity* property of matrix multiplication and the *distributive* property of matrix multiplication and addition.

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, two of the many ways to compute $ABCD$ are $(A\,(BC))\,D = (AB)\,(CD)$: 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

[1] It is hard to appreciate this fully without looking at books written before the linear algebra revolution. Look, for example, at an old British book on the "theory of determinants" or the book *Mathematical Physics*, by Morse and Feshbach. This is great stuff, which is easier to say in modern terminology.

have the same shape (number of rows and columns) as B, and that m_1 and m_2 are multipliers (numbers). Then $A(m_1B_1 + m_2B_2) = m_1(AB_1) + m_2(AB_2)$. This works with more than two B matrices, and with matrices on the right and left, such as

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

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

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

This may be said in a more abstract way. If B is a random matrix and A and C are fixed, not random, then (We use $\mathrm{E}[\cdot]$ to represent expected value.)

$$\mathrm{E}[ABC] \;=\; A\,\mathrm{E}[B]\; C \;. \tag{3}$$

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.

Householder reflections give a nice illustration of matrix distributivity and associativity. Suppose $x \in \mathbb{R}^d$ with $\|x\|_2^2 = x^t x = \sum_i x_i^2 = 1$. The matrix

$$V = I - 2xx^t$$

represents reflection about the plane normal to the vector x. To see this, let $y \in \mathbb{R}^d$ be an arbitrary vector and calculate

$$Vy = y - 2\left(x^t y\right) x \;. \tag{4}$$

Note, in doing this calculation we used distributivity, associativity, and the fact that the 1×1 matrix $x^t y$ is a number that commutes with matrices. If y is perpendicular to x, then $Vy = y$. Otherwise, the formula (4) reverses the sign of the inner product of x and y. That is the reflection, Vy is the "mirror image" of y through the plane perpendicular to x. In particular, $\|Vy\|_2 = \|y\|_2$, which makes the transformation V *orthogonal*. We can see directly, which is the point of this paragraph, that V is an orthogonal matrix, by showing that $VV^t = I$. The interesting part of the calculation, for us here, is when $(xx^t)(xx^t)$ becomes $x\,(x^t x)\, x^t$, which is associativity of matrix multiplication. The inner part on the right is $x^t x = 1$

$$\begin{aligned}
VV^t &= \left(I - 2xx^t\right)\left(I - 2xx^t\right)^t \\
&= \left(I - 2xx^t\right)\left(I - 2xx^t\right) \\
&= I - 2xx^t I - I2xx^t + 4\left(xx^t\right)\left(xx^t\right) \\
&= I - 4xx^t + 4x\left(x^t x\right)x^t \\
&= I - 4xx^t + 4xx^t \\
&= I \;.
\end{aligned}$$

Matrix multiplication is not commutative: $AB \neq BA$ in general. The matrix transpose and matrix inverse reverse the order of matrix multiplication: $(AB)^t = (B^t)(A^t)$, and $(AB)^{-1} = \left(B^{-1}\right)\left(A^{-1}\right)$. For matrix inverses, there is a simple interpretation. The operation A^{-1} undoes the operation of A. The product AB means "first do B, then do A. To undo this, you first undo A, then undo B. In matrix form, this is $B^{-1}A^{-1}$. A *left inverse* of A is a square matrix B so that $BA = I$. A matrix may have a left inverse without being square or having a right inverse. A theorem of linear algebra states that if A is square B is a left inverse, then B is also a right inverse, which means that $AB = I$. Even though $BA \neq AB$ most of the time, if $BA = I$, then $AB = I$. An $m \times n$ matrix has m rows and n columns. If $m > n$, then the matrix is "tall and thin". If $m < n$m then it is "short and fat". A tall and thin matrix can have a left inverse but not a right inverse. A short and fat matrix can have a right inverse but not a left inverse.

We illustrate matrix algebra in probability by finding transformation rules for the mean and covariance of multivariate random variables under linear transformations. Suppose X is a d component random variable, and $Y = AX$. It is not necessary here for A to be invertible or square. The mean of X is the d component vector given either in matrix/vector form as $\mu_X = \mathrm{E}[X]$, or in component form as $\mu_{X,j} = \mathrm{E}[X_j]$. The expected value of Y is

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

We may take A out of the expectation because of the linearity of matrix/vector multiplication.

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

$$C_{X,jk} \;=\; \mathrm{E}[(X_j - \mu_{X,j})\,(X_k - \mu_{X,k})] \;.$$

The diagonal entries of C_X are the variances of the components of X:

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

Now consider the $d \times d$ matrix $B(X) = (X - \mu_X)\,(X - \mu_X)^t$. The (j,k) entry of B is $(X_j - \mu_{X,j})\,(X_k - \mu_{X,k})$. Therefore the (j,k) entry of C_X is the expected value of $B(X)_{jk}$. This proves the matrix formula

$$C_X \;=\; \mathrm{E}[B(X)] = \mathrm{E}\Big[(X - \mu_X)\,(X - \mu_X)^t\Big] \;. \tag{5}$$

The linearity formula (3), and associativity, give the transformation law for covariances under linear transformations. If $Y = AX$, then

$$\begin{aligned}
C_Y &= \mathrm{E}\Big[\left(Y-\mu_Y\right)\left(Y-\mu_Y\right)^t\Big] \\
&= \mathrm{E}\Big[\left(AX-A\mu_X\right)\left(AX-A\mu_X\right)^t\Big] && (Y = AX \text{ transformations}) \\
&= \mathrm{E}\Big[\left\{A\left(X-\mu_X\right)\right\}\left\{A\left(X-\mu_X\right)\right\}^t\Big] && (\text{factor out } A) \\
&= \mathrm{E}\Big[\Big\{A\left(X-\mu_X\right)\Big\}\Big\{\left(X-\mu_X\right)^t A^t\Big\}\Big] && (\text{transpose product rule}) \\
&= \mathrm{E}\Big[A\big\{\left(X-\mu_X\right)\left(X-\mu_X\right)^t\big\}A^t\Big] && (\text{associativity}) \\
&= A\,\mathrm{E}\Big[\left(X-\mu_X\right)\left(X-\mu_X\right)^t\Big]\;A^t && (\text{linearity formula (3)}) \\
C_Y &= AC_XA^t\ . && (6)
\end{aligned}$$

As an example, suppose the components of X are independent standard normals. This means that $X_j \sim \mathcal{N}(0,1)$, and X_j is independent of X_k if $j \neq k$. The covariance matrix of X has $\sigma^2_{X_j} = 1$ in the diagonal and $\mathrm{cov}(X_j, X_k) = 0$ on the off diagonals. This is the identity matrix; $C_X = I$. Let A be the lower triangular matrix

$$A = \begin{pmatrix} 1 & 0 & & \dots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 1 & 1 & \cdots & & 1 \end{pmatrix}$$

This is the matrix that takes *partial sums.* If $Y = AX$, then $Y_1 = X_1$, $Y_2 = X_1 + X_2$, and $Y_k = X_1 + \cdots + X_k$. The covariance C_Y can be found from the general formula (6) and direct calculation: $C_Y = AC_XA^t = AA^t$,

$$\begin{aligned}
C_Y &= \begin{pmatrix} 1 & 0 & & \dots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 1 & 1 & \cdots & & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & & \cdots & 1 \\ 0 & 1 & 1 & \cdots & 1 \\ \vdots & 0 & \ddots & & \vdots \\ 0 & & \cdots & 0 & 1 \end{pmatrix} \\
&= \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 2 & 2 & \cdots & 2 \\ 1 & 2 & 3 & 3 & 3 \\ \vdots & & & & \vdots \\ 1 & 2 & 3 & \cdots & d \end{pmatrix}
\end{aligned}$$

The general formula is $C_{Y,jk} = \min(j,k)$.

We can verify this formula directly as follows. If $j = k$, then we have

$$C_{Y,jj} = \mathrm{var}\left(\sum_{i=1}^{j} X_i\right) = j\ ,$$

Because this is the sum of j independent random variables with variance 1. If $j < k$ then $Y_k = Y_j + X_{j+1} + \cdots + X_k$.

$$\begin{aligned} \mathrm{cov}(Y_j, Y_k) &= \mathrm{cov}(Y_j, Y_j) + \mathrm{cov}(Y_j, X_{j+1}) + \cdots + \mathrm{cov}(Y_j, X_k) \\ &= j + 0 + \cdots + 0 \ . \end{aligned}$$

All the covariances after the first on the right are equal to zero because Y_j is independent of X_i for $i > j$.

Sometimes it is natural that the "components" of X themselves are vectors with more than one component. For example, X_j might refer to measurements made at time j. If more than one measurement is made at time j, then X_j would have more than one component. We use the following terminology. If X_j has m_j components, with $m_j > 1$, we say that X is a *block vector* with m_j being the block size of block X_j. If

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

we say that X has d blocks. The number of components altogether is $M = m_1 + \cdots + m_d$.

Linear transformations on block vectors are represented by block matrices:

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1d} \\ A_{21} & A_{22} & & A_{2d} \\ \vdots & & \ddots & \vdots \\ A_{d1} & A_{d2} & \cdots & A_{dd} \end{pmatrix} \tag{7}$$

This would transform a block vector X with block sizes m_k to a block vector Y with block sizes n_j Matrix "entry" A_{jk} has size $n_j \times m_k$. All the matrices on row j have n_j scalar rows. All the matrices on column k have m_k scalar columns. The overall size of A is $N \times M$, where $N = \sum n_j$, and $M = \sum m_k$. The number of scalar rows of A is the number of scalar rows in the first matrix row, which is m_1, plus the number in the second matrix row, which is m_2, and so on. If $Y = AX$, then Y is a block vector with block sizes n_j. The block matrix/block vector product may be written

$$Y_j = \sum_{k=1}^{d} A_{jk} X_k \ .$$

This formula has the same form as the one for ordinary matrix/vector multiplication, except that the scalar components X_k are replaced by multi-component vectors, and the scalar matrix elements are now small matrices.

Two block matrices are compatible for multiplication if all the row and column numbers match: the number of scalar columns in matrix column k of

A must be the same as the number of scalar rows in the matrix row of B. The result is

$$(AB)_{jl} = \sum_k A_{jk}B_{kl} \; .$$

You multiply the matrices by multiplying and adding individual blocks. The difference is that the individual matrix products need not commute: $A_{jk}B_{kl} \neq B_{kl}A_{jk}$. In fact, $B_{kl}A_{jk}$ need not make sense.

1.3 Principal component analysis

In linear algebra you learn about eigenvalues and eigenvectors, and about singular values and singular vectors. In probability, this subject is called *principal component* analysis, or *PCA*. The extra feature in probability is that distinct principal components are uncorrelated to each other. Principal component analysis decomposes a multivariate random vector into a sum of uncorrelated random vectors.

Suppose C is a symmetric $d \times d$ matrix. Then C has d real eigenvalues and orthonormal eigenvectors:

$$Cv_j = \lambda_j v_j \;\; , \quad v_j^t v_k = 0 \; , \text{ if } j \neq k \;\; , \quad v_j^t v_j = \|v_j\|_2^2 = 1$$

The eigenvectors can be assembled into an eigenvector matrix

$$V = \begin{pmatrix} | & | & \cdots & | \\ v_1 & v_2 & \cdots & v_d \\ | & | & \cdots & | \end{pmatrix} \; .$$

The columns of V are the eigenvectors of C. You can check that the eigenvalue relations may be stated in matrix form as

$$\begin{aligned} CV &= C \begin{pmatrix} | & | & \cdots & | \\ v_1 & v_2 & \cdots & v_d \\ | & | & \cdots & | \end{pmatrix} \\ &= \begin{pmatrix} | & | & \cdots & | \\ \lambda_1 v_1 & \lambda_2 v_2 & \cdots & \lambda_d v_d \\ | & | & \cdots & | \end{pmatrix} \\ &= \begin{pmatrix} | & | & \cdots & | \\ v_1 & v_2 & \cdots & v_d \\ | & | & \cdots & | \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & \lambda_d \end{pmatrix} \\ &= V\Lambda \; , \end{aligned}$$

where Λ is the diagonal eigenvalue matrix on the right. The orthogonality relations are equivalent to the matrix relation $V^tV = I$. This implies also that $VV^t = I$. The eigenvalue decomposition can be written in several equivalent ways. Starting with the above $CV = V\Lambda$, we can get either $C = V\Lambda V^t$, or $\Lambda = V^t CV$.

Let $x \in \mathbb{R}^d$ be a vector, and define $y = V^t x$. The component y_j of y is given by $y_j = v_j^t x$. This means that y_j is the component of x in the direction v_j, and x has the PCA representation $x = \sum_j y_j v_j$. The formula $C = V\Lambda V^t$ has the following interpretation. If you want to calculate Cx, first compute $y = V^t x$, which is the same as representing x in terms of the eigenvectors v_j. Then multiply y_j by λ_j, which corresponds to $\Lambda y = \Lambda V^t x$. Then re-assemble $Cx = \sum \lambda_j y_j v_j$, which is the same as $V\Lambda y = V\Lambda V^t x$.

Now suppose X is a d component random variable with mean zero and covariance C. The eigenvectors v_j are not random, but the components $Y_j = v_j^t X$ are. The terminology is uncertain, but either the vectors v_j or the components Y_j are principal components. The expression

$$X = \sum_{j=1}^{d} Y_j v_j \tag{8}$$

represents X as a sum of principal components. There are several easy to derive the formula

$$\mathrm{cov}(Y_j, Y_k) = v_j^t C v_k \ .$$

This shows that Y_j and Y_k are uncorrelated if $j \neq k$. It also shows that λ_j is the variance of Y_j. A typical value of Y_j will be on the order of $\sqrt{\lambda_j}$. We can arrange the eigenvalues in decreasing order; $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. In many practical applications the eigenvalues are *graded*, which means they decrease rapidly as j increases. In this case the Y_j are also likely to be decreasing rapidly. That suggests that the representation (8) of X in terms of principal components is likely to be *efficient* in the sense that the first few terms give a good approximation to the whole sum.

The *singular value decomposition*, or *SVD*, is PCA for non-symmetric matrices. Suppose A is an $m \times n$ matrix. The SVD of A consists of two orthonormal bases and a collection of non-negative stretch factors. The $v_j \in \mathbb{R}^n$, for $j = 1, \ldots, n$, are the *right singular vectors* of A. The $u_k \in \mathbb{R}^m$, for $m = 1, \ldots, m$, are the *right singular vectors* of A. They are an orthonormal basis for the *column space* of A, which is the subspace of $\mathbb{R}^m$ spanned by the columns of A. The stretch factors, σ_j, are *singular values.* These satisfy the relations $Av_j = \sigma_j u_j$. By convention the singular values are listed in decreasing order, $\sigma_1 \geq \sigma_2 \geq \cdots$. The singular vectors are organized into matrices, V and U, whose columns are the v_j and u_j respectively.

There are different conventions about how to treat the fact that A is not square. One is to have either U or V be rectangular. If A is a tall thin matrix ($m > n$, more rows than columns), then we could say that there are n left and right singular vectors, so V is $n \times n$ and U is $m \times n$. This makes Σ, a matrix with the σ_j on the diagonal, also $n \times n$. The singular vector and value relationships are equivalent to the matrix equation $AV = U\Sigma$, or to $A = U\Sigma V^t$. The other convention would be to make U a square matrix by adding orthonormal columns that span the subspace of $\mathbb{R}^m$ that is perpendicular to the column space. In this convention, U is an $m \times m$ orthogonal matrix, V is an $n \times n$ orthogonal matrix, and Σ is $m \times n$, with all zeros except for singular values on the diagonal.

There are two forms of PCA, eigenvalues and eigenvectors for symmetric matrices, singular vectors and singular values for non-symmetric matrices. These are related. If $A = U\Sigma V^t$ and $H = A^t A$, then H is a symmetric matrix, and

$$\begin{aligned} H &= \left(U\Sigma V^t\right)^t \left(U\Sigma V^t\right) \\ &= \left(V\Sigma^t U^t\right)\left(U\Sigma V^t\right) \\ &= V\Sigma^t \left(U^t U\right)\Sigma V^t \\ &= V\Sigma^t \Sigma V^t \\ H &= V\Lambda V^t \quad , \Lambda = \Sigma^t\Sigma \ . \end{aligned}$$

The eigenvalues of $A^t A$ are σ_j^2. The corresponding eigenvectors are the right singular vectors v_j. Similarly, the left singular vectors u_j are the eigenvectors of AA^t. The eigenvalues are the same, almost. If A is not square, then one of $A^t A$ or AA^t has more eigenvalues. The extra eigenvalues are all zero. The non-zero eigenvalues are all of the form σ_j^2 for some j.

Here is one of the many uses of PCA in practice. It often happens that the eigenvalues (for symmetric H) or the singular vectors (for general A) are strongly *graded*. That means that they decrease quickly from one to the next. This means that H or A can be accurately represented by a sum containing just a few principal components

$$A \approx \sum_{j=1}^{r} \sigma_j u_j v_j^t \ .$$

For example, the 500×500 covariance matrix of the stocks in the S&P 500 index is reasonably well represented by $r = 10$ "market factors".

There are some things eigenvalues and eigenvectors can do that singular values and singular vectors cannot do. One is computing a function of a matrix. The eigenvectors of H^2 and H are the same. The eigenvalues of H^2 are λ_j^2, the eigenvalues of H^{-1} are λ_j^{-1}. The PCA of H^2 or H^{-1} are almost the same as the PCE of H. The expression A^2 may not make sense, but even if it does, the singular vectors of A^2 are not the singular vectors of A, and the singular values of A^2 are not functions of the eigenvalues of A. The PCA of A^2 can be very different from the PCA of A.

1.4 Gaussian probability density

This section gives the formula for the multivariate Gaussian probability density function. There are two "parameters", μ and H, where $\mu \in \mathbb{R}^d$ is the *mean*, and H is a symmetric positive definite $d \times d$ matrix called the *precision.* A multivariate random variable, X, is *multivariate normal* if its probability density function (PDF) is a multivariate normal density.

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

We will see that if X is normal (short for "multivariate normal"), then

$$\mu = \mathrm{E}[X] \;, \tag{10}$$

and

$$H^{-1} = \mathrm{cov}(X) \;. \tag{11}$$

The covariance matrix is called Σ, or C, or C_X, or C_{XX}. If X is Gaussian, the distribution of X is completely determined by its mean and covariance matrix. If $V(x)$ is any function, then $E[V(X)]$ is a function of μ and C. There are many explicit formulas of this kind. We write

$$X \sim \mathcal{N}(\mu, C) \;\;, \text{ or } \;\; X \sim \mathcal{N}(\mu, H^{-1}) \;,$$

if X is normal with mean μ and covariance $C = H^{-1}$.

This section explains five properties of the multivariate normal:

1. *Linear functions of Gaussians are Gaussian.* If X is Gaussian and $Y = AX + b$, then Y is Gaussian. (*Warning*: There is a technical catch.)

2. *Conditioned Gaussians are Gaussian.* Suppose X is a block vector with components X_1 and X_2. If X is a multivariate normal, then the distribution of X_1, conditioned on knowing the value $X_2 = x_2$, is Gaussian.

3. *Marginals of Gaussians are Gaussian.* Suppose X is a block vector with components X_1 and X_2. If X is a multivariate normal, then the distribution of X_1 (ignoring X_2) is Gaussian.

4. *Uncorrelated Gaussians are independent.* Suppose X is a block vector with components X_1 and X_2. If $\mathrm{cov}(X_1, X_2) = 0$ then X_1 and X_2 are independent.

5. Suppose $X = (X_1, X_2)$ in block form. Suppose that the marginal of X_1 is Gaussian (more simply, suppose X_1 is Gaussian). Suppose that the conditional distribution of X_2, given that $X_1 = x_1$ is Gaussian with conditional mean $\mu_2(x_1) = Ax_1 + b$, and precision H_{22} that does not depend on x_1. Then X is Gaussian. The case $A = 0$ is important. In this case the conditional distribution of X_2 does not depend on X_1. That means X_2 is independent of X_1. If X_1 and X_2 are independent Gaussians, then (X_1, X_2) is jointly Gaussian.

Each of these is a theorem about the multivariate PDF formula (9), so we use the PDF formula to prove them But once they are proven, we try as much as possible to think about Gaussians using these properties rather than the PDF that underlies them.

We start with a few simple remarks about the Gaussian PDF. In terms of the covariance, it is

$$u(x) \;=\; \frac{1}{\sqrt{(2\pi)^d \mathrm{det}(C)}}\, e^{-(x-\mu)^t C^{-1}(x-\mu)/2} \;. \tag{12}$$

If $d = 1$, then C is a 1×1 matrix, whose only entry is $\sigma^2 = \mathrm{var}(X)$. Then $X \sim \mathcal{N}(\mu, \sigma^2)$ if X has density

$$u(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} \;. \tag{13}$$

Perhaps the most complicated aspect of the PDF (9) is the *prefactor*

$$\frac{\sqrt{\mathrm{det}(H)}}{(2\pi)^{d/2}} \;.$$

If $f(x)$ is any non-negative function, with a finite integral, and

$$Z = \int f(x)\, dx \;,$$

then

$$u(x) = \frac{1}{Z} f(x) \tag{14}$$

is a probability density. The constant $\frac{1}{Z}$ is the *normalization constant* or *prefactor*. It is common to have a formula for $f(x)$ but not for Z. Even if there is a formula for Z, it may be so complicated that we try to use it as little as possible. In the Gaussian case,

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

The normalization constant is given by the integral

$$\int e^{-(x-\mu)^t H(x-\mu)/2}\, dx = \frac{(2\pi)^{d/2}}{\sqrt{\mathrm{det}(H)}} \;.$$

Then shortened form

$$u(x) = \frac{1}{Z} e^{-(x-\mu)^t H(x-\mu)/2} \;.$$

may be easier to work with than the fully written out version (9).

Mathematicians often use Z to mean "some normalization factor", so (14) is understood to say: "$u(x)$ is equal to $f(x)$ up to some normalization factor". For example, we might write

$$\begin{aligned} u_1(x) &= \frac{1}{Z}\frac{1}{1+x^2} \\ u_2(x) &= \frac{1}{Z} e^{-|x|} \;. \end{aligned}$$

We understand that the Z in the first formula is not the same as the Z in the second one.

It is often useful to treat the Gaussian density, up to a normalization factor, as the exponential of a quadratic form plus a linear form. A *quadratic form*, written $Q(x)$, is a function of the form $Q(x) = \sum_{jk} h_{jk} x_j x_k$. This is to say, a function is a quadratic form if it can be written in the form $Q(x) = x^t H x$. A positive definite quadratic form is one that satisfies $Q(x) > 0$ if $x \neq 0$, which is the same as H being a positive definite matrix. It often happens that a quadratic form is specified in some other way, such as

$$Q(x) = x_1^2 + (x_2 - x_1)^2 + \cdots + (x_{d-1} - x_d)^2 + x_d^2 \; . \tag{15}$$

This is obviously positive definite, but it does not exhibit H explicitly. We can find the entries of H by multiplying out:

$$\begin{aligned} Q(x) &= x_1^2 + \left[x_2^2 - 2x_1x_2 + x_1^2\right] + \cdots + \left[x_2^2 - 2x_1x_2 + x_1^2\right] + x_d^2 \\ &= 2x_1^2 - 2x_1x_2 + x_2^2 - 2x_2x_3 + \cdots + 2x_{d-1}^2 - 2x_{d-1}x_d + 2x_d^2 \end{aligned}$$

The matrix form $x^t H x$ multiplies out to be

$$h_{11}x_1^2 + 2h_{12}x_1x_2 + \cdots + 2h_{1d}x_1x_d + h_{22}x_2^2 + 2h_{23}x_2x_3 + \cdots$$

The off diagonal terms get a factor of 2 because, for example, $h_{12}x_1x_2 = h_{21}x_2x_1$. Comparing expressions (for example, $-2x_1x_2 = 2h_{12}x_1x_2$) shows that the H for our Q is

$$h_{jj} = \begin{cases} 2 & \text{if } j = k \; , \\ -1 & \text{if } j = k \pm 1 \; , \\ 0 & \text{if } |j - k| > 1 \; . \end{cases}$$

This is

$$H = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & -1 & 2 & & 0 \\ \vdots & & \ddots & \ddots & -1 \\ 0 & \cdots & & -1 & 2 \end{pmatrix} \tag{16}$$

This is a *tridiagonal* matrix, with all diagonal entries equal to 2 and main off diagonal entries equal to -1. It may not be obvious from the matrix form (16) that H is positive definite. But it is obvious from quadratic form representation (15). A *linear form* is a function of the form $x^t b$. A general quadratic polynomial is the sum of a quadratic form, a linear form and a constant.

We show that a Gaussian probability density is any PDF that can be written as the exponential of a quadratic polynomial:

$$u(x) = \frac{1}{Z} e^{-\frac{1}{2}Q(x) + b^t x + c} \; . \tag{17}$$

Of course, we can set $c = 0$ by changing Z, or we can set $Z = 1$ by changing c. For example, the PDF of a *Gaussian random walk* with d steps is

$$u(x) = \frac{1}{Z_d} e^{-\frac{1}{2}x_1^2 + (x_2 - x_1)^2 + \cdots + (x_{d-1} - x_d)^2} \ .$$

We will find it convenient to derive this form for the PDE and then, if necessary, to identify H. We show this is Gaussian by putting it in the form (9). We saw that there is an H so that $Q(x) = x^t Hx$. We just need to identify μ, which we do by "completing the square".

$$\begin{aligned} (x-\mu)^t H(x-\mu) + c &= x^t Hx + x^t b \\ x^t Hx - 2x^t H\mu + \mu^t H\mu + c &= x^t Hx + x^t b \ . \end{aligned}$$

We find μ by matching the linear parts from both sides, $-2x^t H\mu = x^t b$. This is supposed to hold for every x, so $H\mu = b$, which is $\mu = -\frac{1}{2}H^{-1}b$. We don't bother finding the constant because it can be absorbed into Z in the end.

Property 1, nonsingular A**.** We leave out the b at first, so $Y = AX$ and $X = A^{-1}Y$. We distinguish the parameters for two PDF functions $X \sim u(x)$ and $Y \sim v(y)$ by using subscripts μ_X, H_{XX}, μ_Y, and H_{YY}. We take $\mu_X = 0$ at first. The linear change of variable formula (1) (with A instead of M) gives the PDF of Y as

$$v(y) = \frac{1}{Z_Y} e^{-\left(A^{-1}y\right)^t H_{XX} \left(A^{-1}y\right)/2}$$

We compute the exponent first, then the normalization constant. We write A^{-t} for $\left(A^{-1}\right)^t$. The notation makes sense because $\left(A^{-1}\right)^t = \left(A^t\right)^{-1}$.

$$\begin{aligned} \left(A^{-1}y\right)^t H_{XX} \left(A^{-1}y\right) &= y^t A^{-t} H_{XX} A^{-1} y \\ &= y^t H_{YY} y \ , \end{aligned}$$

with

$$H_{YY} = A^{-t} H_{XX} A^{-1} \ . \tag{18}$$

This shows that

$$v(y) = \frac{1}{Z_Y} e^{-y^t \, H_{YY} \, y/2} \ .$$

This shows that the probability density of y also has the form of a multivariate normal. Once you know this, there is a simpler way to derive the relation (18) between the precision matrices, and the new normalization constant Z_Y. If you take away the assumptions $b = 0$ and $\mu_X = 0$, a similar but slightly longer calculation shows that

$$v(y) = \frac{1}{Z_Y} e^{-(y-\mu_Y)^t H_{YY} (y-\mu_Y)/2} \ ,$$

with the same H relation (18) and the intuitively obvious

$$\mu_Y = A\mu_X + b \ .$$

We will come back later, twice, to discuss what can happen if A as singular.

Property 2. We can think of H has having block structure corresponding to the block structure of X:

$$\begin{aligned} x^t H x &= \begin{pmatrix} x_1^t & x_2^t \end{pmatrix} \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= x_1^t H_{11} x_1 + 2x_1^t H_{12} x_2^t + x_2^t H_{22} x_2 \ \ . \end{aligned} \tag{19}$$

Because H is symmetric, the off diagonal blocks satisfy the relation

$$H_{12} = H_{21}^t \ .$$

Therefore $x_1^t H_{12} x_2 = x_2^t H_{21} x_1$. These two terms have been combined in (19). If x_1 and x_2 have n_1 and n_2 components respectively, then H_{12} is $n_1 \times n_2$ and H_{21} is $n_2 \times n_1$.

In general, if $u(x_1, x_2)$ is the joint PDF, then the conditional distribution of X_1 given $X_2 = x_2$ is

$$u(x_1 | x_2) = \frac{1}{Z(x_2)} u(x_1, x_2) \ .$$

If you don't care about normalization constants, the conditional density formula for x_1 and joint density of (x_1, x_2) are the same. The normalization constant can depend on x_2, though it will turn out to be independent of x_2 in the Gaussian case. We plug in (19) to get

$$u(x_1 | x_2) = \frac{1}{Z(x_2)} e^{-\frac{1}{2} x_1^t H_{11} x_1 - x^t H_{12} x_2} \ . \tag{20}$$

The term $x_2^t H_{22} x_2$ was not left out. It was "absorbed into the constant" $Z(x_2)$. As a function of x_1 it is indeed a constant.

The conditional PDF formula (20) makes it "obvious" that the conditional x_1 distribution is Gaussian. That's because the exponent is a quadratic function of x_1. We can put it in the specific form (9) by completing the square. We want the exponent in the form $(x_1 - \mu_{X_1}(x_2))^t \, H_{11} \, (x_1 - \mu_{X_1}(x_2)) + w(x_2)$. The leftover term $w(x_2)$ will be absorbed into the normalization constant. Multiplying it out gives

$$(x_1 - \mu_{X_1}(x_2))^t \, H_{11} \, (x_1 - \mu_{X_1}(x_2)) = x_1^t H_{11} x_1 - 2x_1^t H_{11} \mu_{X_1}(x_2) + \cdots \ .$$

This matches (19), up to stuff that depends only on x_2 if we match the term that is linear in x_1. That leads to

$$x_1^t H_{11} x_2 = -x_1^t H_{11} \mu_{X_1}(x_2) \ .$$

This is supposed to be true for every x_1, which gives

$$\begin{aligned} H_{11} x_2 &= -H_{11} \mu_{X_1}(x_2) \\ \mu_{X_1}(x_2) &= H_{11}^{-1} x_2 \ . \end{aligned} \tag{21}$$

This proves property 2, but there is a simpler way to derive the formula for the conditional mean.

Property 4, part 1. Suppose the off-diagonal terms in the precision matrix are zero: $H_{12} = 0$ and $H_{21} = 0$. Then

$$\begin{aligned} u(x_1, x_2) &= \frac{1}{Z} e^{-\frac{1}{2}\left[(x_1-\mu_1)^t H_{11}(x_1-\mu_1) + (x_2-\mu_2)^t H_{22}(x_2-\mu_2)\right]} \\ &= \frac{1}{Z_1} e^{-\frac{1}{2}(x_1-\mu_1)^t H_{11}(x_1-\mu_1)} \; \frac{1}{Z_2} e^{-\frac{1}{2}(x_2-\mu_2)^t H_{22}(x_2-\mu_2)} \\ u(x_1, x_2) &= u_1(x_1)\, u_2(x_2) \; . \end{aligned} \tag{22}$$

This shows that if the off diagonal entries in the precision matrix vanish, then the corresponding block components are independent. We still need to show that the off diagonal blocks of the precision matrix are zero if and only if the off diagonal blocks of the covariance matrix are zero. That is a bit of linear algebra we will do soon.

Property 3. Suppose at first that $\mu = (\mu_1, \mu_2) = 0$. If $X = (X_1, X_2)$, the marginal distribution of X_1 is

$$u_1(x_1) = \int u(x_1, x_2)\, dx_2 \; .$$

We will see that X_1 is Gaussian by showing that

$$u_1(x_1) = \frac{1}{Z_1} e^{-\frac{1}{2} x_1^t \widetilde{H}_{11} x_1} \; .$$

We do this by using properties 1 and 4, and some block linear algebra. Define a block linear transformation of the form

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} I & 0 \\ -K & I \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \; .$$

This is the block matrix way of writing the pair of equations

$$\begin{aligned} y_1 &= x_1 \\ y_2 &= x_2 - K x_1 \; . \end{aligned}$$

The idea is to choose the *feedback* matrix (or *gain* matrix) K so that Y_1 is uncorrelated with Y_2. That implies that Y_1 is independent of Y_2. This, in turn, implies that Y_1 is Gaussian. But $Y_1 = X_1$, so there we are.

Soon, but not now, we will do this calculation with covariances. Now we do it with precision matrices. The precision matrix for (Y_1, Y_2) is given by the transformation formula (18). We need the formula for A^{-1}, which is just as it would be if H_{XX} were a 2×2 scalar matrix rather than a block matrix:

$$A^{-1} = \begin{pmatrix} I & 0 \\ K & I \end{pmatrix} \; , \text{ because } \begin{pmatrix} I & 0 \\ K & I \end{pmatrix} \begin{pmatrix} I & 0 \\ -K & I \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} \; .$$

Therefore

$$\begin{aligned}
H_{YY} &= \begin{pmatrix} I & K^t \\ 0 & I \end{pmatrix} \begin{pmatrix} H_{11} & H_{12} \\ H_{12}^t & H_{22} \end{pmatrix} \begin{pmatrix} I & 0 \\ K & I \end{pmatrix} \\
&= \begin{pmatrix} H_{11} + K^t H_{12}^t & H_{12} + K^t H_{22} \\ H_{12}^t & H_{22} \end{pmatrix} \begin{pmatrix} I & 0 \\ K & I \end{pmatrix} \\
&= \begin{pmatrix} H_{11} + K^t H_{12}^t + H_{12}K + K^t H_{22} K & H_{12} + K^t H_{22} \\ H_{12}^t + H_{22}K & H_{22} \end{pmatrix}
\end{aligned}$$

We want the off diagonal blocks to be zero, which gives

$$\begin{aligned}
0 &= H_{12} + K^t H_{22} \\
K^t &= -H_{12} H_{22}^{-1} \\
K &= H_{22}^{-1} H_{12}^t \ .
\end{aligned} \tag{23}$$

We now write H_{YY}, with this K, as

$$H_{YY} = \begin{pmatrix} \widetilde{H}_{11} & 0 \\ 0 & H_{22} \end{pmatrix} \ .$$

The H_{22} block is the same as before. Substituting (23) gives

$$\begin{aligned}
\widetilde{H}_{11} &= H_{11} - H_{12} H_{22}^{-1} H_{12}^t - -H_{12} H_{22}^{-1} H_{12}^t + H_{12} H_{22}^{-1} H_{22} H_{22}^{-1} H_{12}^t \\
\widetilde{H}_{11} &= H_{11} - H_{12} H_{22}^{-1} H_{12}^t \ .
\end{aligned} \tag{24}$$

This calculation shows that $Y_1 = X_1$ is Gaussian with precision matrix given by (24). It also shows how convenient block matrix notation can be.

Property 1, short and fat A**.** Suppose $Y = AX$ where A is an $n \times m$ matrix with $n < m$ but A having full rank $\text{rank}(A) = n$. In this case there are fewer Y variables than X variables. The mapping is *onto* if A has rank n, which means that for any y there is at least one x with $y = Ax$. For $n < m$ there is a hyperplane of x values that satisfy $y = Ax$, the hyperplane having dimension $m - n$. This includes the case $n = 1$, in which case there is only one Y component. In that case, we might write the single row of A as a^t, and write $Y = a^t X$. We would call Y a *linear functional* of x. The case $n = 2$ involves two linear functionals, which we can write $Y_1 = a_1^t X$ and $Y_2 = a_2^t X$. These are (as we will see) jointly Gaussian.

We use the above properties to show that Y is multivariate normal. The singular value decomposition of A is $A = U \Sigma V^t$, where U is non-singular $n \times n$ and V is non-singular $m \times m$. Then Σ is an $n \times m$ matrix with block form $\begin{pmatrix} \widetilde{\Sigma} & 0 \end{pmatrix}$, where $\widetilde{\Sigma}$ is a square $n \times n$ matrix that is invertible because it is diagonal with singular values $\sigma_j > 0$ on the diagonal (because A has full rank n). Since X is Gaussian and V^t is non-singular, property 1 implies that $Z = V^t X$ is Gaussian. Think of Z as a block vector with Z_1 having the first m components and Z_2 having the remaining $n-m$ components. Property 3 tells us that Z_1 is Gaussian. Finally, property 1 tells us that $Y = U\widetilde{\Sigma} Z_1$ is Gaussian, because $U\widetilde{\Sigma}$ is non-singular. This reasoning may be written out as

$$Y = U \begin{pmatrix} \widetilde{\Sigma} & 0 \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = U \begin{pmatrix} \widetilde{\Sigma} & 0 \end{pmatrix} \begin{pmatrix} & \\ & V^t & \\ & \end{pmatrix} \begin{pmatrix} \\ X \\ \\ \end{pmatrix} .$$

We say a little about the case $n > m$ below.

Property 5. The joint density of $X = (X_1, X_2)$ is the product of the marginal density of X_1 and the conditional density of X_2, given X_1. In a formula,

$$u(x_1, x_2) = u_1(x_1)\, u_{2,1}(x_2 | x_1) \ . \tag{25}$$

By hypothesis X_1 is Gaussian, so

$$u_1(x_1) = \frac{1}{Z_1} e^{-\frac{1}{2}(x_1-\mu_1)^t H_{11}(x_1-\mu_1)} \ .$$

Also by hypothesis, the conditional distribution of X_2 for given x_1 is

$$u_{12}(x_2|x_1) = \frac{1}{Z_2} e^{-\frac{1}{2}(x_2-[Ax_1+b])^t H_{22}(x_2-[Ax_1+b])} \ .$$

It is important that even though the distribution of X_2 depends on x_1, the normalization constant Z_2 is independent of x_1. In fact, the formula is

$$Z_2 = \frac{(2\pi)^{m/2}}{\sqrt{\det(H_{22})}} \ ,$$

where m is the number of scalar components of X_2. We assumed that the conditional precision (later, the conditional covariance) of X_2 is independent of x_1. With these expressions, the joint density becomes

$$u(x_1, x_2) = \frac{1}{Z_1 Z_2} e^{-\frac{1}{2}R(x_1,x_2)} \ ,$$

where the exponent is

$$R(x_1, x_2) = (x_1 - \mu_1)^t \, H_{11} \, (x_1 - \mu_1) + (x_2 - [Ax_1 + b])^t \, H_{22} \, (x_2 - [Ax_1 + b]) \ .$$

If you multiply this out, you will see that it is a quadratic (plus linear plus constant) in (x_1, x_2). You will also get explicit formulas for the blocks of the resulting precision matrix.

We have been assuming that the normalization constant in (9) is correct. One way to verify the normalization constant is to use the eigenvalue and eigenvector decomposition $H = V\Lambda V^t$. We already used the fact the eigenvector matrix V has $\det(V) = 1$. We can represent X in terms of the eigenvectors

$$X = \sum_{i=1}^{d} Y_i v_i \ .$$

The expansion coefficients are $Y_i = v_i^t X$. This is expressed in matrix vector form as $Y = V^t X$. The v_i, or the Y_i, or both are called principal components. The PDf of Y has precision matrix $H_{YY} = V^t H_{XX} V = \Lambda$, so

$$\begin{aligned} v(y) &= \frac{1}{Z} e^{-\frac{1}{2} y^t \Lambda y} \\ &= \frac{1}{Z} e^{-\lambda_1 y_1^2/2} \cdots e^{-\lambda_d y_d^2/2} \end{aligned} \tag{26}$$

We use the “well known” (those who don’t know it should look it up) formula

$$\int_{-\infty}^{\infty} e^{-z^2/2}\, dz = \sqrt{2\pi}$$

More generally, the substitution $z = \sqrt{\lambda} y$, and $dz = \sqrt{\lambda} dy$ makes this into

$$\int_{-\infty}^{\infty} e^{-\lambda y^2/2}\, dy = \sqrt{\frac{2\pi}{\lambda}}$$

The normalization constant in (26) is

$$\begin{aligned} Z &= \int_{\mathbb{R}^d} e^{-\frac{1}{2} v^t \Lambda v} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\lambda_1 y_1^2/2} \cdots e^{-\lambda_d y_d^2/2}\, dy_1 \cdots dy_d v \\ &= \sqrt{\frac{2\pi}{\lambda_1}} \sqrt{\frac{2\pi}{\lambda_2}} \cdots \sqrt{\frac{2\pi}{\lambda_d}} \\ &= \frac{(2\pi)^{d/2}}{\left(\prod_{i=1}^d \lambda_i\right)^{1/2}} \\ &= \frac{(2\pi)^{d/2}}{\sqrt{\det(H)}} \ . \end{aligned}$$

Since $\det(V) = 1$, the general PDF transformation formula (1) implies that this Z is also the normalization constant for $u(x)$.

1.5 Using linear algebra and the covariance matrix

From the point of view of probability it may be more natural to calculate with the covariance matrix than with the precision matrix. If X is a d component random variable, the individual means and covariances are $\mu_j = E[X_j]$, and

$$C_{jk} = \text{cov}(X_j, X_k) = E[(X_j - \mu_j)(X_k - \mu_k)] \ . \tag{27}$$

These are organized into the vector mean and the covariance matrix as $\mu = E[X]$ and

$$C = \text{cov}(X) = E\Big[(X - \mu)(X - \mu)^t\Big] \ . \tag{28}$$

You should check that the (j,k) entry of the matrix (28) is the scalar formula (27). It is clear that the μ parameter in (9) is the mean of a Gaussian. For the remainder of this section, we set $\mu = 0$ to focus on the covariance.

We verify the relation between the covariance and precision matrices using a little linear algebra, the covariance transformation formula (6), and the independence property (4). Using the $H = V\Lambda V^t$ eigenvalue and eigenvector decomposition, we again let Y be the vector of principal component amplitudes, $Y = V^t X$, with $X = VY$. Then

$$\begin{aligned} C_{XX} &= E[XX^t] \\ &= E[VYY^tV^t] \\ &= VE[YY^t]V^t \\ &= VC_{YY}V^t \ . \end{aligned}$$

The components Y_j are independent (because $H_{YY} = \Lambda$ is diagonal), so the off diagonal entries of C_{YY} are zero. The diagonal entries are

$$C_{YY,jj} = \mathrm{var}(Y_j) = E[Y_j^2] \ .$$

The PDF of Y_j is

$$v_j(y_j) = \frac{1}{Z_j} e^{-\frac{1}{2}\lambda_j y_j^2} \ .$$

From this, we have $Z_j = \sqrt{\frac{2\pi}{\lambda_j}}$ and

$$E[Y_j^2] = \sqrt{\frac{\lambda_j}{2\pi}} \int_{-\infty}^{\infty} y_j^2 e^{-\frac{1}{2}\lambda_j y_j^2} \, dy_j = \frac{1}{\lambda_j}$$

The actual evaluation at the end may be done by substituting $\lambda_j y_j^2 = z^2$, which is $z_j = \sqrt{\lambda_j} y_j$. The result is

$$C_{YY} = \Lambda^{-1} \ .$$

This gives

$$C_{XX} = V\Lambda^{-1}V^t = H_{XX}^{-1} \ . \tag{29}$$

This shows that the relation between covariance and precision is $C = H^{-1}$. That is why the PDF formulas (9) and (12) are equivalent.

Property 4, part 2. Suppose X is a block vector of the form

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

The covariance matrix of X has a corresponding block form

$$\begin{aligned} C_{XX} &= E[XX^t] \\ &= E\left[\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\begin{pmatrix} X_1^t & X_2^t \end{pmatrix}\right] \\ &= E\begin{bmatrix} X_1X_1^t & X_1X_2^t \\ X_2X_1^t & X_2X_2^t \end{bmatrix} \\ &= \begin{pmatrix} C_{11} & C_{12} \\ C_{12}^t & C_{22} \end{pmatrix} \ . \end{aligned}$$

The diagonal blocks are

$$C_{11} = \text{cov}(X_1) \ \ , \quad C_{22} = \text{cov}(X_2) \ .$$

The off diagonal blocks are

$$C_{12} = \text{cov}(X_1, X_2) = E[X_1X_2^t] \ \ , \quad C_{21} = \text{cov}(X_2, X_1) = E[X_2X_1^t] = C_{12}^t \ .$$

The covariance matrix is *block diagonal* if the off diagonal blocks vanish: $C_{12} = 0$ and $C_{21} = 0$. We don't say $C_{12} = C_{21} = 0$ because C_{12} and C_{21} have different shapes if X_1 and X_2 have different number of scalar components. The inverse of a block diagonal matrix, if it exists, is block diagonal. Therefore, C_{XX} is block diagonal if and only if H is block diagonal. If X_1 and X_2 are uncorrelated, which is the same as saying $C_{12} = 0$, then H_{XX} is block diagonal, which implies that X_1 and X_2 are independent.

1.6 Generating a multivariate normal, interpreting covariance

Monte Carlo simulation with Gaussians is easy because there are simple algorithms to generate a Gaussian with a specified covariance C. You start with $Z \sim \mathcal{N}(0, I)$, which is the same as d independent standard normals $Z_1, \ldots, Z_d$. A *standard* normal is a scalar Gaussian with mean zero and variance 1. Most programming systems have standard random number generators. In R, the command is

```
Z = rnorm(d)
```

The next step is to find a matrix M so that $X = MZ$ has the desired covariance C. The transformation law (6) in this case is $C_X = MC_ZM^t$. But $C_Z = I$ by construction, so we need M with

$$MM^t = C \ . \tag{30}$$

Such an M would be a kind of square root of C. It is not unique. Even the square root of 4 is not unique, because $2^2 = (-2)^2 = 4$. It is possible to find such an M as long as C is symmetric and positive definite. We will see two distinct ways to do this which 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. Such an L exists as long as C is SPD (symmetric and positive definite). 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). Most programming languages have commands to compute L. In R, it is

```
L = chol(C)
```

R uses `%*%` to represent matrix-vector or matrix-matrix multiplication. So the following code will produce and use N independent Gaussians with covariance C

```
L = chol(C)
for ( i in (1:n)){
   Z = nrand(d)
   X = L \%*\% Z
        ...
      (use X)
        ...
 }
```

Different calls to `nrand()` produce independent Z vectors, so the X vectors are also independent. The most expensive single operation is the Cholesky step. Leaving it out of the loop means that we pay this overhead just once even though we generate a large number of random vectors, X.

Consider as an example the two dimensional case with $\mu = 0$. Here, we want X_1 and X_2 that are jointly normal. We specify $\text{var}(X_1) = \sigma_1^2$, $\text{var}(X_2) = \sigma_2^2$, and the *correlation coefficient*

$$\rho_{12} \;=\; \text{corr}(X_1, X_2) \;=\; \frac{\text{cov}(X_1, X_2)}{\sigma_1\sigma_2} \;=\; \frac{\text{E}(X_1X_2)}{\sigma_1\sigma_2} \; .$$

The corresponding scalar covariance is $C_{12} \;=\; \text{cov}(X_1, X_2) \;=\; \rho_{12}\sigma_1\sigma_2$. The target covariance matrix is

$$C = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

In this case, the Cholesky factor is (check this)

$$L \;=\; \begin{pmatrix} \sigma_1 & 0 \\ \rho_{12}\sigma_2 & \sqrt{1-\rho_{12}^2}\sigma_2 \end{pmatrix} \; . \tag{31}$$

The formula $X = LZ$ becomes

$$X_1 = \sigma_1 Z_1 \tag{32}$$

$$X_2 = \rho_{12}\sigma_2 Z_1 + \sqrt{1-\rho_{12}^2}\,\sigma_2 Z_2 \; . \tag{33}$$

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

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

which is the desired answer, too. The scalar covariance is is also correct:

$$\mathrm{cov}(X_1,X_2) \;=\; \mathrm{E}[X_1X_2] \;=\; \mathrm{E}\left[\sigma_1 Z_1 \rho_{12}\sigma_2 Z_1\right] \;=\; \rho_{12}\sigma_1\sigma_2\,\mathrm{E}\big[Z_1^2\big] \;=\; \rho_{12}\sigma_1\sigma_2 \; .$$

The formulas (32) and (33) have natural interpretations. First, X_1 is determined by a standard normsl *factor*, which we call Z_1. The scaling σ_1 gives X_1 the desired variance. The expression (33) has the same factor Z_1 scales by the desired correlation, ρ_{12}, and then by σ_2 so that X_2 will eventually have the desired variance. With just $\rho_{12}\sigma_1 Z_1$, our X_2 would have variance $\rho_{12}^2\sigma_2^2$, which is too small. We add in an independent contribution with variance $\sigma_2^2(1-\rho_{12}^2)$ to get the desired variance for X_2. The second factor, Z_2, contributes only to X_2.

We could have turned the formulas (32) and (33) around as

$$\begin{aligned} X_1 &= \sqrt{1-\rho_{12}^2}\,\sigma_1 Z_1 \;+\; \rho_{12}\sigma_1 Z_2 \\ X_2 &= \qquad\qquad\qquad\;\; \sigma_2 Z_2 \; . \end{aligned} \tag{34}$$

This starts by building X_2 with the right variance, and then re-using Z_2 to build X_1 with the desired variance and correlation to X_2. It gives $X = MZ$ with

$$M = \begin{pmatrix} \sqrt{1-\rho_{12}^2}\,\sigma_1 & \rho_{12}\sigma_1 Z_2 \\ 0 & \sigma_2 \end{pmatrix} \; .$$

You should check that this M also satisfies $MM^t = C$.

The two approaches to generating X give different pictures of causation. In the first one, X_1 seems to exert an influence on X_2. In the second, it is the reverse. This illustrates an important saying in statistics: Correlation does not imply causation. If we observe that X_1 and X_2 are correlated we do not necessarily know why. Did X_1 happen first and then influence X_2, or maybe X_2 happened first and influenced X_1, or maybe they both were influenced by a factor we did not observe.

2 Linear Gaussian recurrences

Discrete time is measured in discrete time units $n = 0, 1, 2, \ldots$. A discrete time Gaussian *process* is a sequence, $X = X_1, X_2, ...$, so that X is a Gaussian. If each X_n has d components, and if n runs from 1 to T, then X may be thought of as a blocked Gaussian with T blocks of size d. This means that each individual X_n is a d component Gaussian. But X being Gaussian says a lot more. For example, each pair (X_n, X_{n+1}) is a $2d$ component block Gaussian with two blocks of size d. We call X the *path*, and X_n the *value* of the path at time n.

A two term linear Gaussian recurrence is a relation of the form

$$X_{n+1} \;=\; AX_n \;+\; BZ_n \; . \tag{35}$$

The $Z_n \sim \mathcal{N}(0, I)$ are independent m component normals, with $m \leq d$. This can be written in many ways. Some economists would write

$$X_{n+1} = AX_n + \epsilon_n \; ,$$

where the *residuals* ϵ are independent $\mathcal{N}(0, C_{\epsilon\epsilon})$. The two forms are equivalent, if we take B with $BB^t = C_{\epsilon\epsilon}$). It seems obvious, and we will soon verify, that if the X_n satisfy a linear Gaussian recurrence, and if X_1 is Gaussian, then the path X is also Gaussian. We will see how to find the big $dT \times dT$ covariance matrix C_{XX} from the covariance matrix of X_1 and B.

Linear Gaussian recurrences are a class of stochastic processes. We think of X_n as the *state of the system* at time n. The dynamical equation (35) says that the state at time $n+1$ depends linearly on the state at time n, but knowing X_n does not determine X_{n+1} completely. There is "noise", which is random input Z_n or ϵ_n, which is independent of the path up to time n. The random input may also be called the *innovation* (economics), or the *shock* (finance, bankers must be easily shocked).

Linear Gaussian recurrences are used to model systems ranging from evolving economic states, to the heaving of the surfaces of stars, to the buffeting of an airplane by turbulent air patterns. More realistic models would have nonlinear dynamics, with $X_n \to AX_n$ replaced by a nonlinear function, and non-Gaussian forcing, possibly with intermittency or fat tails. We discuss Gaussian processes here for two reasons. One is that they are good models for some problems, like star surface motion under some circumstances. The other is that they illustrate many features of more general stochastic evolution systems.

2.1 The probability density

We want to characterize the probability density of the path. As in the general discussion of Gaussians, this can be done in two steps. First we see that the joint distribution is Gaussian, then we identify the parameters using linear algebra.

Suppose X_1 is Gaussian with mean μ_1 and covariance $C_{X_1X_1}$. We work by induction on t. The *path* up to time n is $X_{[1:t]}$. This is the block vector with td components and blocks $X_1, \ldots, X_t$. The induction hypothesis is that $X_{[1:t]}$ is a block Gaussian vector with td components and blocks $X_1, \ldots, X_t$. We suppose this is true and describe the distribution of the longer path $X_{[1:t+1]}$, which may be thought of as a block vector with blocks $X_{[1:t]}$ and X_{t+1}:

$$X_{[1:t+1]} = (X_1, \ldots, X_t, X_{t+1}) = ((X_1, \ldots, X_t), X_{t+1}) = (X_{[1:t]}, X_{t+1}) \; .$$

The conditional distribution of X_{t+1} given $x_{[1,t]}$ is the same as the conditional distribution of X_{t+1} given x_t, which is Gaussian with mean Ax_t and covariance BB^t. Property 5 then implies that the joint distribution is Gaussian. The base case that is needed to start the induction, is that the one state path $X_1 = X_{1:1}$ is Gaussian. The conclusion is that a linear Gaussian recurrence starting with a Gaussian initial state gives a Gaussian path.

2.2 Probability distribution dynamics

Since the path $X_{[1:t]}$ is Gaussian, and X_n is a component of the path if $n \leq t$, we know that X_n is Gaussian (property 3). Let its parameters be $\mu_n = E[X_n]$, and $C_n = \mathrm{cov}(X_n)$. These parameters satisfy recurrence relations that are the key to understanding linear Gaussian dynamics. Taking expectations on both sides of the recurrence relation (35) gives

$$\mu_{n+1} = A\mu_n \; . \tag{36}$$

This says that the recurrence relation for the means is the same as the recurrence relation (35) for the random states if you "turn off the noise" (set Z_n to zero).

For the covariance, it is convenient to combine (35) and (36) into

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

The covariance calculation starts with

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

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

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

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

$$\begin{aligned} C_{n+1} &= \mathrm{E}\Big[\left(A\left(X_n - \mu_n\right)\right)\left(A\left(X_n - \mu_n\right)\right)^t\Big] \;+\; \mathrm{E}\Big[\left(BZ_n\right)\left(BZ_n\right)^t\Big] \\ &= \mathrm{E}\Big[A\Big\{\left(X_n - \mu_n\right)\left(X_n - \mu_n\right)^t\Big\}A^t\Big] \;+\; \mathrm{E}\big[B\left(Z_n Z_n^t\right)B^t\big] \\ &= A\,\mathrm{E}\Big[\left(X_n - \mu_n\right)\left(X_n - \mu_n\right)^t\Big]\;A^t \;+\; B\,\mathrm{E}\big[Z_n Z_n^t B^t\big]\;B^t \\ C_{n+1} &= AC_nA^t \;+\; BB^t \; . \end{aligned} \tag{37}$$

The recurrence relations (36) and (37) determine the distribution of X_{n+1} in terms of the distribution of X_n.

A *forward equation* is an equation that determines the PDF of X_{n+1} in terms of the PDF of X_n. The equations (36) and (37) play the role of a forward equation for a two term linear Gaussian recurrence.

2.3 Higher order recurrence relations, the Markov property

It is common to consider recurrence relations with more than two terms, or more than one lag. A k lag relation has the form

$$X_{n+1} \;=\; A_0 X_n \,+\, A_1 X_{n-1} \,+\, \cdots \,+\, A_{k-1} X_{n-k+1} \,+\, B Z_n \;. \tag{38}$$

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* reformulates a multi-lag system into the form of a two term recurrence relation (35). We start with the k lag system (38) and create an equivalent one lag system. The state for the one lag system, which we call $\widetilde{X}_n$, is a block vector whose blocks are the k lagged states

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

If the states X_n have d components, then $\widetilde{X}_n$ has 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 relation in the expanded state formulation involves block matrices $\widetilde{A}$ and $\widetilde{B}$:

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

In more detail, this is

$$\begin{pmatrix} X_{n+1} \\ X_n \\ \vdots \\ X_{n-k+2} \end{pmatrix} = \begin{pmatrix} A_0 & A_1 & \cdots & & A_{k-1} \\ I & 0 & \cdots & & 0 \\ 0 & I & \cdots & & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & & I & 0 \end{pmatrix} \begin{pmatrix} X_n \\ X_{n-1} \\ \vdots \\ X_{n-k+1} \end{pmatrix} + \begin{pmatrix} B \\ 0 \\ \vdots \\ 0 \end{pmatrix} Z_n \;. \tag{39}$$

The top rows of $\widetilde{A}$ and $\widetilde{B}$ encode the original lagged dynamics (38). The second row of $\widetilde{A}$ equates X_n on the left with X_n on the right, and so on. The matrix $\widetilde{A}$ is the *companion matrix* of the recurrence relation (38).

We will see in subsection 2.5 that the stability of a recurrence relation (35) is determined by the eigenvalues of A. For the case $d = 1$, you might know that the stability of the recurrence relation (38) 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 (38), then the covariance matrix, $\widetilde{C}_n = \text{cov}(\widetilde{X}_n)$, 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 successive states in a one lag system (35) satisfy the *Markov property*: The distribution of X_{t+1} conditional on knowing X_t is the same as the distribution of X_{t+1} knowing the whole path $X_{[1:t]}$. Roughly speaking, the present is all the information about the past that is relevant for predicting the future. A sequence of states that satisfy the Markov property is a *Markov chain*. The k lag system (38) does not satisfy the Markov property if $k > 1$. Knowing X_{t-1} and X_t allows more accurate predictions of X_{t+1} than are possible with just X_t.

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. For linear Gaussian recurrences, the expanded state $\widetilde{X}_n$ has all the information about the past that is relevant for predicting the future. We know this because the $\widetilde{X}_n$ satisfy a one lag recurrence. There are many stochastic processes that are not linear Gaussian recurrences. State space expansion is a common way to study a general process using the theory of Markov processes.

2.4 Unit roots, the borderline case

The borderline case in Subsection 2.5 is eigenvalues on the unit circle in the complex plane. Morally, such a system is mildly unstable. The simplest example is $d = 1$, and $A = 1$ (a 1×1 matrix), and $B = 1$, which gives

$$X_{n+1} = X_n + Z_n \; . \tag{40}$$

From this it follows that $\mu_n = \mu_1$ for all n. We calculate that

$$\text{var}(X_{n+1}) = \text{var}(X_n) + 1 \; .$$

If $X_0 = 0$, then $\sigma_n^2 = \text{var}(X_n) = n$. Clearly $\sigma_n^2 \to \infty$ as $n \to \infty$. This is a mild instability. There is no limiting distribution as $n \to \infty$, but the variance grows linearly rather than exponentially.

More generally, if A has an eigenvalue with $|\lambda| = 1$, then either $\lambda = \pm 1$ or λ is somewhere on the unit circle, so $\overline{\lambda}$ is also an eigenvalue. In any of these cases, unless the problem has a degeneracy, the variance grows linearly with n (reasoning omitted). Problems like this are common in applications, either simple random walks or more complicated processes with random walk components. In finance, *co-integration* is the phenomenon that $|\lambda_j| \leq 1$ for all j and there is at least one eigenvalue with $|\lambda_j| < 1$. Discovering co-integration is not easy and can be rewarding.

2.5 Large time behavior and stability

We often want to understand things about a stochastic process that do not depend on the initial state. We may start observing a system only after it has been "running" so long that its initial state is forgotten. *Large time behavior* is the behavior of X_n as $n \to \infty$. The stochastic process (35) 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 (36) and (37). The limiting distribution depends on the following limits:

$$\mu = \lim_{n\to\infty} \mu_n \tag{41}$$

$$C = \lim_{n\to\infty} C_n \tag{42}$$

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

Mathematicians say that something is *morally true* if it is true in almost any real problem, and if the only situations in which it is not true seem contrived or unnatural. In that sense, it is morally true that a one lag linear Gaussian recurrence has a statistical steady state if and only if the noise free dynamics ((35) with $Z_n = 0$) is strongly stable in the sense that $X_n \to 0$ as $n \to \infty$. This theorem, (stochastic steady state exists) $\Longleftrightarrow$ (strong linear stability) is true without exceptions if the noise matrix B is square and has rank d. The derivation and proof are simpler if A has d linearly independent eigenvectors, which is to say that it has no non-trivial Jordan blocks. We discuss this case first, then come back to the situations where B is rectangular or A has Jordan blocks.

The eigenvalues and eigenvectors of A satisfy $Ar_j = \lambda_j r_j$, for $j = 1, \ldots, d$. The λ_j and r_j may be complex. The notation r_j is for *right* eigenvector. There are also *left* eigenvectors, which are row vectors that satisfy $l_j A = \lambda_j l_j$. The eigenvectors form a basis of $\mathbb{C}^d$, which implies that the eigenvector matrix

$$R = \begin{pmatrix} | & | & & | \\ r_1 & r_2 & \cdots & r_d \\ | & | & & | \end{pmatrix}$$

is non-singular. The eigenvalue and eigenvector relationships are expressed in matrix form as

$$AR = R\Lambda \; . \tag{43}$$

We define $L = R^{-1}$ and multiply (43) by L on both sides, which gives

$$LA = \Lambda L \; . \tag{44}$$

This means that the rows of L are left eigenvectors of A:

$$L = \begin{pmatrix} - & l_1 & - \\ - & l_2 & - \\ & \vdots & \\ - & l_d & - \end{pmatrix}$$

The matrix relation $LR = I$ is equivalent to the *bi-orthogonality* relations

$$l_j r_k = \left\{ \begin{array}{ll} 1 & \text{if } j = k \; , \\ 0 & \text{if } j \neq k \; . \end{array} \right.$$

If the means have an expansion in the right eigenvector basis as $\mu_n = \sum_{j=1}^{d} m_{n,j} r_j$, then $m_{n,j} = l_j \mu_n$. The dynamics (36) imply that $m_{n+1,j} = \lambda_j m_{n,j}$. Therefore

$$m_{n,j} \;=\; \lambda_j^n m_{0,j} \; . \tag{45}$$

The limit (41) 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 unique representations $\mu_n = \sum_{j=1}^{d} m_{n,j} r_j$. The dynamics (36) implies that $m_{n+1,j} = \lambda_j m_{n,j}$. This implies that

$$m_{n,j} \;=\; \lambda_j^n m_{0,j} \; . \tag{46}$$

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 $n \to \infty$ regardless 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 (42) 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

$$\|A^n\| \;\leq\; c\, a^n \; , \tag{47}$$

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 (42) exists by writing C as a convergent infinite sum. To simplify notation, write R for BB^t. Suppose C_0 is given, then (37) gives $C_1 = AC_0A^t + R$. Using (37) 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

$$C_n = A^nC_0\left(A^n\right)^t + \sum_{k=0}^{n-1} A^kR\left(A^k\right)^t\ . \tag{48}$$

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

$$C = \sum_{k=0}^{\infty} A^kR\left(A^k\right)^t\ . \tag{49}$$

For the first term, note that (47) and properties of matrix norms imply that[2]

$$\left\|A^nC_0\left(A^n\right)^t\right\| \le \left(ca^n\right)\left\|C_0\right\|\left(ca^n\right) = ca^{2n}\left\|C_0\right\|\ .$$

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:

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

Properties of norms bound this by a geometric series:

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

You can find C without summing the infinite series (49). Since the limit (42) exists, you can take the limit on both sides of (37), which gives

$$C - ACA^t = BB^t\ . \tag{50}$$

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

Subsection 2.6 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, (50) 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 (37) and the steady state equation (50) 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 (49), or by solving (50). 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 (49) diverges.

2.6 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 (37). These are related to the eigenvalues and eigenvectors of A.

The covariance recurrence relation (37)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

$$C_{n+1} \;=\; AC_nA^t \; . \tag{51}$$

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, (37) 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 (51) 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:

$$\begin{aligned} AS_{jk}A^t \;&=\; A\left(r_j r_k^t + r_k r_j^t\right) A^t \\ &=\; A\left(r_j r_k^t\right) A^t \;+\; A\left(r_k r_j^t\right) A^t \\ &=\; \left(Ar_j\right)\left(Ar_k\right)^t \;+\; \left(Ar_k\right)\left(Ar_j\right)^t \\ &=\; \left(\lambda_j r_j\right)\left(\lambda_k r_k\right)^t \;+\; \left(\lambda_k r_k\right)\left(\lambda_j r_j\right)^t \\ &=\; \lambda_j\lambda_j\left(r_j r_k^t + r_k r_j^t\right) \\ &=\; \mu_{jk}S_{jk} \;. \end{aligned}$$

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 (37) 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 (50) 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 (50) 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.

3 Estimation, filtering, prediction

Gaussian models are often used for estimation and filtering. You have a quantity, X, whose value you do not know. You have some data, Y, whose value depends partly on X. The *estimation* problem is to say something about X, given Y. Suppose X_n are the successive states in a discrete time linear one lag Gaussian process. Suppose Y_n is an *observation* of X_n. This means that Y_n depends in some way on X_n. The *filtering* problem is to say something about X_n from the observation path $Y_{[1:n]}$. The *prediction* problem is to say something about X_{n+k} for $k > 0$, given the observation path up to time n.

There are two common points of view regarding estimation and filtering, the *frequentist* and the *Bayesian* (after Bayes) views. A frequentist either does not regard X not as being random, or does not feel it is appropriate to create a model of the random value X might have. Scientists who estimate the speed of light, c, or other physical constants from data often take a frequentist point of view. A frequentist constructs an *estimator*, $\widehat{Y}$, which is the best guess of the value of X given the data. For example, suppose a science team) measures the travel time of a laser beam to the moon and back. If the round trip distance is D and the times are T_n, $\ldots$, T_N, then the team could use the average travel time $\overline{T} = \frac{1}{N}\sum T_k$ and estimate $\widehat{c} = D/\overline{T}$. Or they could average the individual measured (with measurement error) travel speeds $c_k = R/T_k$ and average those: $\widehat{c} = \frac{1}{N}\sum_k c_k$. Both of these estimators are functions of the data $Y = (T_1, \ldots, T_N)$, but they are not the same. Statistical theory gives insight which one might be better in which circumstances. Theory also gives *error bars*, which are indications of the size of $|\widehat{c} - c|$. This error is random not because c is random, but because Y is random.

A Bayesian team regards X as a random variable with a known (or, more properly, assumed) PDF, $u(x)$. They also assume a conditional PDF $L(y|x)$, that describes the conditional PDF of Y given X. This is often a physical model of the noisy observation process. The notation L comes from *likelihood*, which is a statisticians' term for probability when it is not a function of the random variable. The joint density of (X, Y) is $u(x)L(y|x)$. The conditional density of X, given the observation $Y = y$, is called the *posterior* distribution $u(x|y)$, which is given by Bayes' rule of conditional probability

$$u(x|y) = \frac{1}{Z(y)} u(x)L(y|x) \; . \tag{52}$$

The data are useful if the uncertainty in X after knowing the data is less than the uncertainty in X in the prior. The normalization constant $Z(y)$ is determined, as usual, by the requirement that $u(x|y)$ should be a PDF as a function of x. This gives

$$Z(y) = \int u(x)L(y|x)\, dx \; .$$

In practice, the normalization constant is hard to determine. Bayesians instead use Monte Carlo sampling methods to create samples of the posterior distribution.

One advantage of the frequentist approach is that it is easier to do and easier to understand. The computations involved are usually optimization (maximum likelihood), solving nonlinear equations (generalized method of moments), etc.

There is a single reported answer, $\widehat{x}$. Bayesian statistic is harder to do. You need to describe the posterior distribution, often by finding many samples of it. Not only is sampling harder to do than optimization, but it is harder to explain to the customer that the information coming from the data is contained in a list of samples.

These issues and tradeoffs are different in the Gaussian case than they are in general. One reason is that a Gaussian distribution is completely described by its mean and covariance matrix. There is no need to sample to represent the posterior. Computing the posterior mean and covariance usually boils down to numerical linear algebra, which is "easy" (given our computational infrastructure) except for very large problems. The posterior mean may well be the frequentist maximum likelihood estimate, which makes Bayesian and frequentist results nearly the same.

3.1 Partial information and conditional distributions

We did this earlier when we verified property 2 earlier. We do it again here using covariances, and we put the result in more the *feedback* form that will be useful in several specific cases. The general form here will be specialized in several ways, and gives a common framework for the specific examples.

Suppose $X = (X_1^t, X_2^t)^t$ is a block column vector. If the overall mean is zero, the block covariance matrix is

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} = \begin{pmatrix} E[X_1 X_1^t] & E[X_1 X_2^t] \\ E[X^2 X_1^t] & E[X_2 X_2^t] \end{pmatrix} \; .$$

We want to describe the information you get about X_2 by knowing X_1. A frequentist might give an estimator of X_2 that is a function of X_1. This would be written $\widehat{X}_2(X_1)$. A Bayesian might try to describe the conditional probability density $u(x_2|x_1)$. As we have already seen, and will see again in a slightly different way, these two things are more or less the same in the Gaussian setting.

It is natural, or will soon seem so, to make a prediction that is a linear function of the data:

$$\widehat{X}_2(X_1) = KX_1 \; , \tag{53}$$

where K, which is called the *feedback*, or *gain*, is a matrix of the appropriate dimensions. The *residual* is the prediction error

$$\epsilon = X_2 - \widehat{X}_2 = X_2 - KX_1 \; ,$$

or

$$X_2 = KX_1 + \epsilon \; . \tag{54}$$

We will choose K so that the residual is uncorrelated with the data. Property 4 implies that ϵ is independent of X_1. Then (54) implies that X_2 is equal the estimator plus a residual that is independent of X_1. A frequentist might argue that the independence of the residual from the data makes the estimator (53) the best possible for this situation. A Bayesian might use this to give the conditional distribution of X_2 a Gaussian with mean KX_1 and covariance $\mathrm{cov}(\epsilon)$.

The actual algebra is simpler than the philosophy. The covariance of ϵ with X_1 is

$$\begin{aligned}\mathrm{cov}(\epsilon, X_1) &= E[\epsilon X_1^t] \\ &= E[(X_2 - KX_1)X_1^t] \\ &= C_{21} - KC_{11} \ .\end{aligned}$$

We set the covariance to zero and solve for K, which yields

$$K = C_{21}C_{11}^{-1} \ . \tag{55}$$

The remaining uncertainty in X_2, after the data X_1 and the prediction (53) is (use (55) and $C_{12} = C_{21}^t$)

$$\begin{aligned}\mathrm{cov}(\epsilon) &= E[\epsilon\epsilon^t] \\ &= E[(X_2 - KX_1)(X_2 - KX_1)^t] \\ &= E[(X_2 - KX_1)(X_2^t - X_1^tK^t] \\ &= E[X_2X_2^t] - E[X_2X_1^t]K^t - KE[X_1X_2^t] + KE[X_1X_1^t]K^t \\ &= C_{22} - C_{21}C_{11}^{-1}C_{21}^t - C_{21}C_{11}^{-1}C_{21}^t + C_{21}C_{11}^{-1}C_{11}C_{11}^{-1}C_{21}^t \\ \mathrm{cov}(\epsilon) &= C_{\epsilon\epsilon} = C_{22} - C_{21}C_{11}^{-1}C_{21}^t \ . \end{aligned} \tag{56}$$

This formula has a natural interpretation. C_{22} is the uncertainty in X_2 before learning the data. When you subtract from X_2 the prediction using X_1, you reduce the uncertainty. The term $C_{21}C_{11}^{-1}C_{21}^t$ quantifies how much the uncertainty is reduced. Of course, if $C_{21} = 0$ then X_1 is independent of X_2, so knowing X_1 does not give any information on X_2. In that case $K = 0$, which means the optimal prediction ignores X_1. Using X_1 in a non-trivial way would increase the uncertainty in X_2, not decrease it.

3.2 Using a noisy observation

Suppose $X \sim \mathcal{N}(0, C)$ and $Y = BX + W$, with $W \sim \mathcal{N}(0, R)$, and W independent of X, is a *noisy observation* of X. Some applications have X with many components and Y with just a few, so B already looses information. In that case, even a noise free observation, which corresponds to $R = 0$ would not allow us to predict X exactly. Other applications have a lot of noisy data, so the *observation matrix*, B, could be tall and thin. We want to find the optimal estimator of the form $\widehat{X} = KY$ and the covariance matrix describing the remaining uncertainty. It is possible to work this out "from scratch" (not using Subsection 3.1). But we instead reformulate this prediction problem in the general form of Subsection 3.1) and use the solution there.

We create a block vector with components $X_1 = Y$, which is the data, and $X_2 = X$, which is to be predicted. The necessary matrices are

$$\begin{aligned} C_{11} &= \mathrm{cov}(Y) \\ &= E[YY^t] \\ &= E[(BX+W)(BX+W)^t] \\ &= BCB^t + R \; . \end{aligned}$$

The omitted terms such as $BE[XW^t]$ vanish because the observation noise is independent of X. The other matrix is

$$C_{21} = E[XY^t] = E[X(X^tB^t + W^t)] = CB^t \; .$$

The general formula (55) becomes

$$K = CB^t(BCB^t + R)^{-1} \; .$$

The prediction error is $\epsilon = X - KY$, and the covariance of the prediction error is

$$C_{\epsilon\epsilon} = C - CB^t(BCB^t + R)^{-1}BC \; .$$