Multivariate Gaussian

We say that a random variable $X \in \mathbb{R}^D$ is a **Gaussian random vector** if there exists a vector $\boldsymbol{\mu} \in \mathbb{R}^D$ and a symmetric positive definite matrix $\boldsymbol{R}$ such that its density can be written as

$$f_X(\boldsymbol{x}) = \frac{1}{(2\pi)^{D/2}\sqrt{\det(\boldsymbol{R})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

The vector $\boldsymbol{\mu}$ is the mean of this distribution, and $\boldsymbol{R}$ is the covariance:

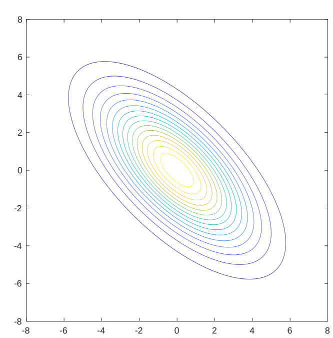
$$\boldsymbol{\mu} = \mathrm{E}[X], \quad \boldsymbol{R} = \mathrm{E}[(X-\boldsymbol{\mu})(X-\boldsymbol{\mu})^{\mathrm{T}}].$$

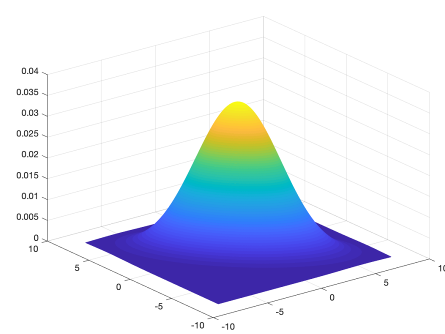
We will denote this as

$$X \sim \mathrm{Normal}(\boldsymbol{\mu}, \boldsymbol{R}).$$

The geometry of the density reflects the eigenstructure of $\boldsymbol{R}$ — the level-surfaces of the density are ellipsoids with the eigenvectors of $\boldsymbol{R}$ as axes, and radii are proportional to the eigenvalues of $\boldsymbol{R}$.

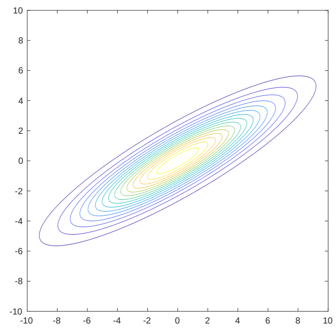
$$\begin{aligned}\boldsymbol{\mu} &= \begin{bmatrix}0\\0\end{bmatrix}\\ \boldsymbol{R} &= \begin{bmatrix}6 & -4\\ -4 & 6\end{bmatrix}\\ &= \frac{1}{2}\begin{bmatrix}-1 & 1\\ 1 & 1\end{bmatrix}\begin{bmatrix}10 & 0\\ 0 & 2\end{bmatrix}\begin{bmatrix}-1 & 1\\ 1 & 1\end{bmatrix}\end{aligned}$$

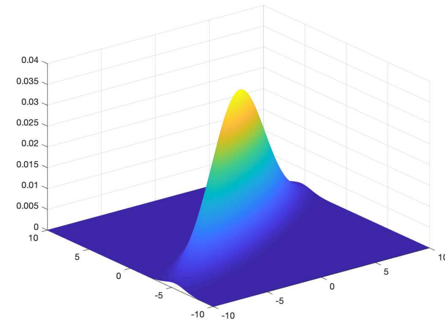




$$
\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}
$$

$$
\begin{aligned}
\boldsymbol{R} &= \begin{bmatrix} 15.25 & 8.23 \\ 8.23 & 5.75 \end{bmatrix} \\
&= \begin{bmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{bmatrix} \begin{bmatrix} 20 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{3}}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} \end{bmatrix}
\end{aligned}
$$





It is clear from its definition that $\boldsymbol{R}$ is symmetric. It is also postive semidefinite[1], as

$$
\begin{aligned}
\boldsymbol{w}^{\mathrm{T}} \boldsymbol{R} \boldsymbol{w} = \boldsymbol{w}^{\mathrm{T}} \operatorname{E}[(X - \boldsymbol{\mu})(X - \boldsymbol{\mu})^{\mathrm{T}}] \boldsymbol{w} &= \operatorname{E}[\boldsymbol{w}^{\mathrm{T}} (X - \boldsymbol{\mu})(X - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{w}] \\
&= \operatorname{E}[|\boldsymbol{w}^{\mathrm{T}} (X - \boldsymbol{\mu})|^2].
\end{aligned}
$$

Since $|\boldsymbol{w}^{\mathrm{T}}(X - \boldsymbol{\mu})|^2$ is always non-negative, its expectation is non-negative as well. So all the eigenvalues are $\lambda_d \geq 0$.

Other facts about the multivariate Gaussian:

1. If $X \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{R})$, then for and $\boldsymbol{w} \in \mathbb{R}^D$, $Y = \boldsymbol{w}^{\mathrm{T}} X$ is a Gaussian (scalar) random variable with

$$
Y \sim \text{Normal}(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{\mu}, \boldsymbol{w}^{\mathrm{T}} \boldsymbol{R} \boldsymbol{w}).
$$

   You will prove this on the homework. Note in particular that this means each entry in X is itself a Gaussian random variable, as taking $\boldsymbol{w}$ as a unit vector with entry i equal to 1 and zero elsewhere, we have

$$
X_i \sim \text{Normal}(\mu_i, R_{ii})
$$

[1] This is not just a property of the multivariate Gaussian; the covariance matrix of any random vector will be symmetric positive semidefinite.

2. If $X \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{R})$ and $\boldsymbol{A}$ is a $M \times D$ matrix, then $Y = \boldsymbol{A}X$ is a Gaussian random vector:
$$Y \sim \text{Normal}(\boldsymbol{A\mu}, \boldsymbol{ARA}^{\mathrm{T}}).$$
That the mean is $\boldsymbol{A\mu}$ follows directly from the linearity of the expectation operator, and that each entry of Y is Gaussian follows directly from our first fact. The only thing we have to check is the expression for the covariance:
$$\begin{aligned}\mathrm{E}[(Y - \boldsymbol{A\mu})(Y - \boldsymbol{A\mu})^{\mathrm{T}}] &= \mathrm{E}[\boldsymbol{A}(X - \boldsymbol{\mu})(X - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}] \\ &= \boldsymbol{A}\,\mathrm{E}[(X - \boldsymbol{\mu})(X - \boldsymbol{\mu})^{\mathrm{T}}]\boldsymbol{A}^{\mathrm{T}} \\ &= \boldsymbol{ARA}^{\mathrm{T}}.\end{aligned}$$

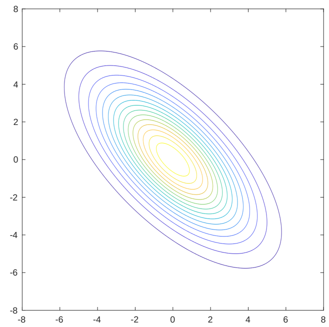
3. If $R_{ij} = 0$, then entries X_i and X_j are independent. You will prove this on the homework.

4. For $X \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{R})$, let
$$\boldsymbol{R} = \boldsymbol{V\Lambda V}^{\mathrm{T}}$$
be the eigenvalue decomposition of $\boldsymbol{R}$. Set
$$\boldsymbol{Z} = \boldsymbol{V}^{\mathrm{T}}X.$$
Then the covariance of Z
$$\mathrm{E}[(Z - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{\mu})(Z - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{\mu})^{\mathrm{T}}] = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{RV} = \boldsymbol{\Lambda},$$
is diagonal. Thus the entries of Z are independent, with variances equal to the eigenvalues:
$$\begin{aligned}Z_1 &\sim \text{Normal}(\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{\mu}, \lambda_1) \\ Z_2 &\sim \text{Normal}(\boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{\mu}, \lambda_2) \\ &\vdots \\ Z_D &\sim \text{Normal}(\boldsymbol{v}_D^{\mathrm{T}}\boldsymbol{\mu}, \lambda_D).\end{aligned}$$

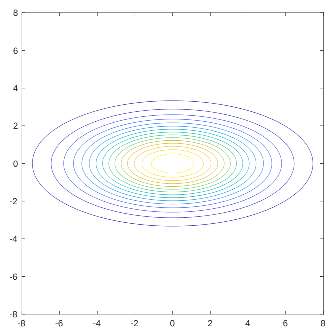
So transforming into the $\boldsymbol{V}$ domain decorrelates the entries of X. This is called the **Karhunen-Loeve** transform.

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{R} = \begin{bmatrix} 6 & -4 \\ -4 & 6 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 10 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}$$



$$\downarrow \quad Z = \boldsymbol{V}^{\mathrm{T}} X$$



$$\boldsymbol{V}^{\mathrm{T}}\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{V}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{V} = \begin{bmatrix} 10 & 0 \\ 0 & 2 \end{bmatrix}$$

Gaussian Estimation

What does observing part of a Gaussian random vector tell us about the part that we do not observe? That is, suppose

$$X \sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{R}),$$

and then we observe the first $1, \ldots, p$ entries of X while entries $p+1, \ldots, D$ stay hidden. We divide X into

$$X = \begin{bmatrix} X_o \\ X_h \end{bmatrix}, \quad \text{then observe} \;\; X_o = \boldsymbol{x}_o.$$

What is the conditional density for $X_h | X_o = \boldsymbol{x}_o$?

It turns out that the conditional density is also Gaussian, just with a different mean and different covariance. To see this, we partition the covariance matrix into 4 parts:

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_o & \boldsymbol{R}_{oh} \\ \boldsymbol{R}_{oh}^{\mathrm{T}} & \boldsymbol{R}_h \end{bmatrix}.$$

The upper left corner contains the $p \times p$ covariance matrix for the random variables that end up being observed, the lower right corner contains the $D - p \times D - p$ covariance matrix for the unobserved random variables, and $\boldsymbol{R}_{oh}$ is the *cross-correlation* matrix, that captures the dependencies between the observed and unobserved random variables.

We can also partition the inverse covariance

$$\boldsymbol{R}^{-1} = \begin{bmatrix} \boldsymbol{B} & \boldsymbol{C} \\ \boldsymbol{C}^{\mathrm{T}} & \boldsymbol{D} \end{bmatrix}.$$

Using the Schur complement (see the Technical Details section below), we can write out these blocks of the inverse as

$$\begin{aligned}
\boldsymbol{D} &= (\boldsymbol{R}_h - \boldsymbol{R}_{oh}^{\mathrm{T}}\boldsymbol{R}_o^{-1}\boldsymbol{R}_{oh})^{-1} \\
\boldsymbol{C} &= -\boldsymbol{R}_o^{-1}\boldsymbol{R}_{oh}\boldsymbol{D} \\
\boldsymbol{B} &= (\text{something})
\end{aligned}$$

We could write down the expression for $\boldsymbol{B}$ if we really wanted to, but it is long and complicated and it ends up that we don't use it. We will use these expressions later, but to ease the notation below, we will stick with $\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$.

We can now compute the conditional density using

$$f_{X_h}(\boldsymbol{x}_h|\boldsymbol{x}_0) = \frac{f_{X_o,X_h}(\boldsymbol{x}_o, \boldsymbol{x}_h)}{f_{X_o}(\boldsymbol{x}_o)}.$$

The numerator is proportional to

$$\begin{aligned}
f_{X_o,X_h}(\boldsymbol{x}_o, \boldsymbol{x}_h) &\propto \exp\left(-\frac{1}{2}\begin{bmatrix}\boldsymbol{x}_o^{\mathrm{T}} & \boldsymbol{x}_h^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\boldsymbol{B} & \boldsymbol{C} \\ \boldsymbol{C}^{\mathrm{T}} & \boldsymbol{D}\end{bmatrix}\begin{bmatrix}\boldsymbol{x}_o \\ \boldsymbol{x}_h\end{bmatrix}\right) \\
&= \exp\left(-\frac{1}{2}\left[\boldsymbol{x}_o^{\mathrm{T}}\boldsymbol{B}\boldsymbol{x}_o + \boldsymbol{x}_o^{\mathrm{T}}\boldsymbol{C}\boldsymbol{x}_h + \boldsymbol{x}_h^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}_o + \boldsymbol{x}_h^{\mathrm{T}}\boldsymbol{D}\boldsymbol{x}_h\right]\right).
\end{aligned}$$

The first term above does not depend on $\boldsymbol{x}_h$, so we can write the conditional density as

$$f_{X_h}(\boldsymbol{x}_h|\boldsymbol{x}_0) = g(\boldsymbol{x}_0)\ \exp\left(-\frac{1}{2}\left[\boldsymbol{x}_o^{\mathrm{T}}\boldsymbol{C}\boldsymbol{x}_h + \boldsymbol{x}_h^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}_o + \boldsymbol{x}_h^{\mathrm{T}}\boldsymbol{D}\boldsymbol{x}_h\right]\right),$$

where $g(\boldsymbol{x}_0)$ is a function that incorporates $1/f_{X_o}(\boldsymbol{x}_o)$ and $\exp(-\boldsymbol{x}_o^{\mathrm{T}}\boldsymbol{B}\boldsymbol{x}_o/2)$ along with some constants. We are not too worried about what g actually is, just that it does not depend on $\boldsymbol{x}_h$.

To show that $X_h|X_o$ is a Gaussian random vector, we need a density that looks like

$$(\text{stuff with no } \boldsymbol{x}_h) \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x}_h - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{K}(\boldsymbol{x}_h - \boldsymbol{\mu})\right).$$

To get our conditional density in this form, we *complete the square* in the exponent. You can easily check that the following relation holds:

$$\begin{aligned}\boldsymbol{x}_o^{\mathrm{T}}\boldsymbol{C}\boldsymbol{x}_h + \boldsymbol{x}_h^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}_o + \boldsymbol{x}_h^{\mathrm{T}}\boldsymbol{D}\boldsymbol{x}_h &= \\ &(\boldsymbol{x}_h + \boldsymbol{D}^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}_o)^{\mathrm{T}}\boldsymbol{D}(\boldsymbol{x}_h + \boldsymbol{D}^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}_o) - \boldsymbol{x}_o^{\mathrm{T}}\boldsymbol{C}\boldsymbol{D}^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}_o.\end{aligned}$$

Again, the last term above does not depend on $\boldsymbol{x}_h$. Thus we can write

$$f_{X_h}(\boldsymbol{x}_h|\boldsymbol{x}_0) = h(\boldsymbol{x}_o)\exp\left(-\frac{1}{2}(\boldsymbol{x}_h + \boldsymbol{D}^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}_o)^{\mathrm{T}}\boldsymbol{D}(\boldsymbol{x}_h + \boldsymbol{D}^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}_o)\right),$$

where $h(\boldsymbol{x}_o) = g(\boldsymbol{x}_o)\exp(\boldsymbol{x}_o^{\mathrm{T}}\boldsymbol{C}\boldsymbol{D}^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}_o/2)$. Plugging in the expressions for $\boldsymbol{C}$ and $\boldsymbol{D}$ above, we see that $X_h|X_o = \boldsymbol{x}_o$ is a Gaussian random vector

$$X_h|X_o = \boldsymbol{x}_o \;\sim\; \mathrm{Normal}(\boldsymbol{R}_{oh}^{\mathrm{T}}\boldsymbol{R}_o^{-1}\boldsymbol{x}_o, \boldsymbol{R}_h - \boldsymbol{R}_{oh}^{\mathrm{T}}\boldsymbol{R}_o^{-1}\boldsymbol{R}_{oh}).$$

So given the observations $X_o = \boldsymbol{x}_o$, our best (MMSE) guess for $\boldsymbol{h}_h$ is the conditional mean:

$$\hat{\boldsymbol{x}}_h = \boldsymbol{R}_{oh}^{\mathrm{T}}\boldsymbol{R}_o^{-1}\boldsymbol{x}_o.$$

The MSE we will incur with this choice is

$$\mathrm{E}[\|\hat{\boldsymbol{x}}_h - X_h\|_2^2 | X_o = \boldsymbol{x}_o] = \mathrm{trace}(\boldsymbol{R}_h - \boldsymbol{R}_{oh}^{\mathrm{T}}\boldsymbol{R}_o^{-1}\boldsymbol{R}_{oh}).$$

It is a fact that $\mathrm{trace}(\boldsymbol{R}_h - \boldsymbol{R}_{oh}^{\mathrm{T}}\boldsymbol{R}_o^{-1}\boldsymbol{R}_{oh}) \leq \mathrm{trace}(\boldsymbol{R}_h)$ (why?), so the observing $X_o = \boldsymbol{x}_o$ also reduces the mean-squared error associated with our best guess.

Notice that for zero-mean Gaussian random variables, $R[i,j] = 0$ if and only if X_i and X_j are independent. Above, this means that if X_o and X_k are independent, we will have $\boldsymbol{R}_{oh} = \boldsymbol{0}$, and the conditional distribution for X_h is no different from its original marginal (exactly as we would expect).

Example: Suppose a Gaussian random vector $X \in \mathbb{R}^2$ has

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{R} = \begin{bmatrix} 6 & -4 \\ -4 & 6 \end{bmatrix}.$$

We will observe $X_1 = x_1$ and see how it affects our outlook on X_2. Before the observation, we have

$$X_2 \sim \text{Normal}(0, 6),$$

so our best estimate at for X_2 is 0, and the mean-square error of this estimate is 6.

Now suppose we observe $X_1 = 4$. How does the distribution for X_2 change? Using the above ($\boldsymbol{R}_o = 6, \boldsymbol{R}_{oh} = -4, \boldsymbol{R}_h = 6$), we have

$$X_2 | X_1 = 4 \ \sim \text{Normal}\left(-\frac{8}{3}, \frac{10}{3}\right).$$

So given $X_1 = 4$, the best estimate for X_2 is now $-8/3$, and the mean-square error for that estimate is $10/3$.

Conditional independence and Gaussian graphical models

Sometimes, asking whether or not two random variables are "independent" is not really getting to the point. For example, let's look at the random vector. Let $Z_1, \ldots, Z_D$ be independent Gaussian random variables with $Z_k \sim \mathrm{Normal}(0,1)$, and $0 < a < 1$, and set

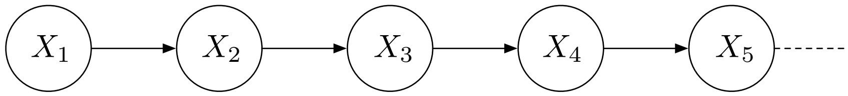
$$\begin{aligned} X_1 &= \sigma\, Z_1, \\ X_2 &= aX_1 + Z_2 \\ X_3 &= aX_2 + Z_3 \\ &\vdots \\ X_D &= aX_{D-1} + Z_D \end{aligned}$$

where $\sigma = (1-a^2)^{-1/2}$ is chosen so that all the X_k have the same variance of $\sigma^2 = (1-a^2)^{-1}$. This is a standard auto-regressive process — the next point in the vector is computed by taking the previous point, multiplying it by a fixed number, then adding an independent perturbation. Here is its covariance matrix:

$$\boldsymbol{R} = \begin{bmatrix} \sigma^2 & a\sigma^2 & a^2\sigma^2 & \cdots & a^{D-1}\sigma^2 \\ a\sigma^2 & \sigma^2 & a\sigma^2 & \cdots & a^{D-2}\sigma^2 \\ \vdots & & \ddots & & \vdots \\ a^{D-1}\sigma^2 & a^{D-2}\sigma^2 & \cdots & & \sigma^2 \end{bmatrix}.$$

None of the random variables are independent of one another (the value of X_1 directly affects the values of X_2 which in turn directly affects the value of X_3 which in turn ... etc.) and this is reflected in the fact that none of the entries in the covariance matrix are equal to zero.

But there is somehow, the "flow" of this process is very naturally described with this graph:



This graph is capturing something subtly different than the statistical dependencies. It says, for example, that if I know that $X_3 = x_3$, then I already have a complete statistical characterization of X_4 ... with $X_3 = x_3$ in hand, it doesn't matter what X_2 was, the conditional distribution (Gaussian with mean ax_3 and variance 1) is set. That is, given X_3, X_2 is **conditionally independent** of X_4.

Conditional independence

How is this notion of conditional independence expressed in the covariance structure? We have seen that It is easy to interpret a zero-valued entry in the covariance matrix: $R_{i,j} = 0$ means X_i and X_j are independent. Conditional independence is expressed in the *inverse covariance* of$X$. Let

$$\boldsymbol{S} = \boldsymbol{R}^{-1}.$$

What does it mean if $S_{i,j} = 0$? The answer is that X_i and X_j are independent given observations of all of the other entries $\{X_k,\ k \neq i, j\}$ in X. To see this, suppose that we partition $\boldsymbol{S}$ the same way we partitioned $\boldsymbol{R}$:

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_o & \boldsymbol{S}_{oh} \\ \boldsymbol{S}_{oh}^{\mathrm{T}} & \boldsymbol{S}_h \end{bmatrix}$$

We have already seen we can use the *Schur complement* to get an expression for $\boldsymbol{S}_h$ in terms of the blocks in $\boldsymbol{R}$:

$$\boldsymbol{S}_h = (\boldsymbol{R}_h - \boldsymbol{R}_{oh}^{\mathrm{T}} \boldsymbol{R}_0^{-1} \boldsymbol{R}_{oh})^{-1}.$$

Notice that this is exactly the inverse covariance of the conditional random vector $X_h|X_o$. Then, consider the particular case when there are two "hidden" entries in X_h, say $X_h = \{X_i, X_j\}$) and the other $d-2$ are in X_o. The off-diagonal terms in $\boldsymbol{S}_h$ above correspond to $S_{i,j}$ and $S_{j,i}$, if these are zero, then $\boldsymbol{S}_h$ is diagonal, and so is $\boldsymbol{S}_h^{-1} = \boldsymbol{R}_{h|o}$. This means that X_i and X_j are **conditionally independent** given observations of $\{X_k,\ k \neq i, j\}$.

Independence and conditional independence

Let $X \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{R})$, and set $\boldsymbol{S} = \boldsymbol{R}^{-1}$. Then

$$R_{i,j} = 0 \quad \Leftrightarrow \quad X_i \text{ and } X_j \text{ are independent},$$

and with $X_{\overline{(i,j)}} = \{X_k, k \neq i, j\}$,

$$S_{i,j} = 0 \quad \Leftrightarrow \quad X_i|X_{\overline{(i,j)}} \text{ and } X_j|X_{\overline{(i,j)}} \text{ are independent.}$$

That $X_i|X_{\overline{(i,j)}}$ and $X_j|X_{\overline{(i,j)}}$ are independent means that we can factor the distribution

$$f_{X_i,X_j}(x_i, x_j|X_{\overline{(i,j)}} = \boldsymbol{v}) = f_{X_i}(x_i|X_{\overline{(i,j)}} = \boldsymbol{v}) \cdot f_{X_j}(x_j|X_{\overline{(i,j)}} = \boldsymbol{v}).$$

Gaussian Graphical Models

Let's return to our example above, where we took $X_k = aX_{k-1} + Z_k$. In this case, it is clear that X_k is conditionally independent of X_i if $|k-i| > 1$. This fact is reflected in the inverse covariance:

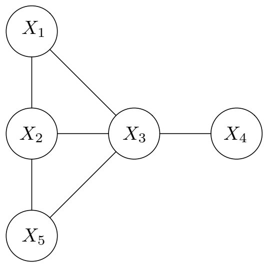
$$\boldsymbol{S} = \boldsymbol{R}^{-1} = \begin{bmatrix} 1 & -a & 0 & \cdots & & 0 \\ -a & (1+a^2) & -a & \cdots & & 0 \\ 0 & -a & (1+a^2) & -a & \cdots & 0 \\ \vdots & & & & & \\ 0 & \cdots & 0 & -a & (1+a^2) & -a \\ 0 & \cdots & & 0 & -a & 1 \end{bmatrix}$$

In general, we can summarize the conditional independence structure using a graph. Each of the nodes of the graph corresponds to an entry X_i, and there is an edge between node i and node j if X_i and X_j are conditionally dependent (i.e. not conditionally independent). Equivalently, if there is not an edge between node i and node j, the corresponding entry of the inverse covariance will be zero.

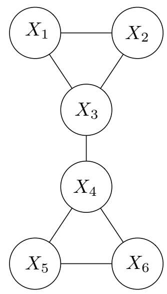
For our particular "chain" example above, we actually only need to observe *one* entry of X_k to make X_{k-1} and X_{k+1} conditionally independent. In fact, observing $X_k = x_k$ divides the remaining parts of the vectors into two groups, $\{X_j,\ j < k\}$ and $\{X_j,\ j > k\}$. Every random variable in the first group will be conditionally independent of every variables in the second group given X_k. Within these groups, the random variables are still not independent given X_k.

This division is represented in the inverse covariance as well. When we observe X_k, the inverse covariance for the $d-1$ random variables that remain "hidden" is block diagonal — the inverse of this matrix (the conditional covariance matrix) will be block diagonal as well.

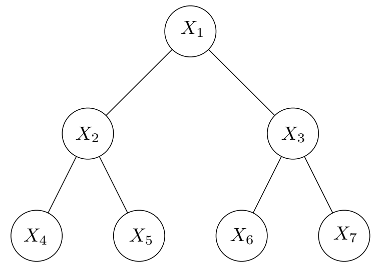
Exercise: For each of the graphs below, indicate the inverse covariance structure



(i)



(ii)



(iii)

Exercise: Suppose that removing vertex X_k separates the graph into two connected components X_{c_1} and X_{c_2} so that there is no path between any vertex in X_{c_1} and X_{c_2}. For example, if we removed X_3 in (ii) above, we could take

$$X_{c_1} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \quad \text{and} \quad X_{c_2} = \begin{bmatrix} X_4 \\ X_5 \\ X_6 \end{bmatrix}.$$

Argue that after observing $X_k = x_k$ at such a node, any $X_i \in X_{c_1}$ and any $X_j \in X_{c_2}$ will be independent; that is

$$f_{X_i,X_j}(x_i, x_j | X_k = x_k) = f_{X_i}(x_i | X_k = x_k) \cdot f_{X_j}(x_j | X_k = x_k).$$

Causality?

The generative equations for our working example above, $X_k = aX_{k-1} + Z_k$, indicated a certain *causal* structure — X_{k-1} is affecting the distribution of X_k in an explicit manner. We emphasized this by drawing the graph structure with connected edges. Unfortunately, this type of causality structure cannot in general be discerned from the covariance matrix (or its inverse). For example, if we take $D = 3$ and $a = 1/2$, we have

$$\boldsymbol{R} = \begin{bmatrix} 2 & 2/\sqrt{2} & 1 \\ 2/\sqrt{2} & 2 & 2/\sqrt{2} \\ 1 & 2/\sqrt{2} & 2 \end{bmatrix}$$

One generating system of equations is

$$\begin{aligned} X_1 &= \sigma Z_1 \\ X_2 &= aX_1 + Z_2 \\ X_3 &= aX_2 + Z_3. \end{aligned}$$

But another one that results in X having exactly the same covariance structure is

$$\begin{aligned} X_1 &= (-0.2881)Z_1 + (-0.7071)Z_2 + (1.1904)Z_3 \\ X_2 &= (0.5219)Z_1 + (1.3144)Z_3 \\ X_3 &= (-0.2881)Z_1 + (0.7071)Z_2 + (1.1904)Z_3. \end{aligned}$$

You can check that $\mathrm{E}[XX^{\mathrm{T}}] = \boldsymbol{R}$ for this second set of equations. And these aren't the only two — there are literally an infinite number of ways we can generate X from Z distributed as above. The point here is that causality is not a property that reveals itself through the covariance matrix.

Technical Details: The Schur Complement

Suppose that $\boldsymbol{M}$ is an invertible matrix broken into four blocks:

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{M}_{11} & \boldsymbol{M}_{12} \\ \boldsymbol{M}_{21} & \boldsymbol{M}_{22} \end{bmatrix}.$$

If $\boldsymbol{M}_{22}$ is invertible, then the inverse of $\boldsymbol{M}$ can be expressed in terms of these blocks using the *Schur complement* of $\boldsymbol{M}$ in $\boldsymbol{M}_{22}$:

$$\boldsymbol{S} = \boldsymbol{M}_{11} - \boldsymbol{M}_{12}\boldsymbol{M}_{22}^{-1}\boldsymbol{M}_{21}.$$

Then,

$$\boldsymbol{M}^{-1} = \begin{bmatrix} \boldsymbol{S}^{-1} & -\boldsymbol{S}^{-1}\boldsymbol{M}_{12}\boldsymbol{M}_{22}^{-1} \\ -\boldsymbol{M}_{22}^{-1}\boldsymbol{M}_{21}\boldsymbol{S}^{-1} & \boldsymbol{M}_{22}^{-1} + \boldsymbol{M}_{22}^{-1}\boldsymbol{M}_{21}\boldsymbol{S}^{-1}\boldsymbol{M}_{12}\boldsymbol{M}_{22}^{-1} \end{bmatrix}$$

Similarly, if $\boldsymbol{M}_{11}$ is invertible, we can do something similar with the Schur complement of $\boldsymbol{M}$ in $\boldsymbol{M}_{11}$:

$$\boldsymbol{M}^{-1} = \begin{bmatrix} \boldsymbol{M}_{11}^{-1} + \boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}\boldsymbol{S}^{-1}\boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1} & -\boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}\boldsymbol{S}^{-1} \\ -\boldsymbol{S}^{-1}\boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1} & \boldsymbol{S}^{-1} \end{bmatrix},$$

where now

$$\boldsymbol{S} = \boldsymbol{M}_{22} - \boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}.$$

These formulas can be checked simply by multiplying out $\boldsymbol{M}^{-1}\boldsymbol{M}$ and seeing that it is $\mathbf{I}$.