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 $Z[k]\sim\text{Normal}(0,1)$, and $0<a<1$, and set

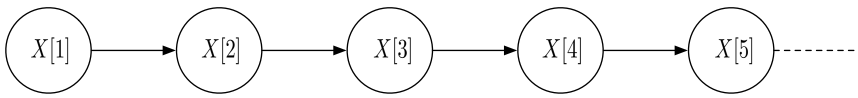
$$\begin{aligned}X[1] &= \sigma\, Z[1],\\ X[2] &= aX[1]+Z[2]\\ X[3] &= aX[2]+Z[3]\\ &\vdots\\ X[d] &= aX[d-1]+Z[d]\end{aligned}$$

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

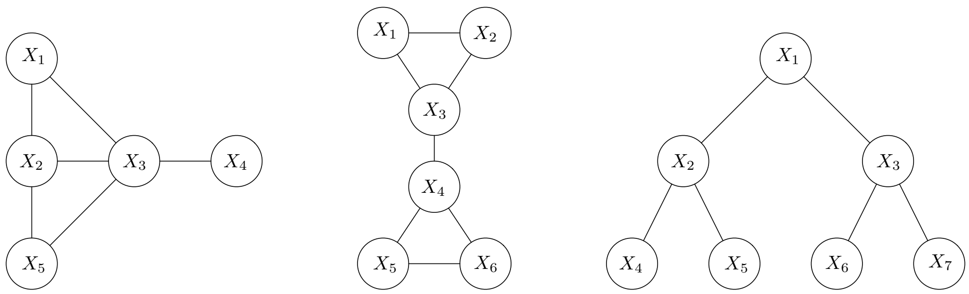
$$\boldsymbol{S} = \boldsymbol{R}^{-1} = \begin{bmatrix} 1 & -a & 0 & \cdots & & 0 \\ -a & (1+a^2) & -a & \cdots & & 0 \\ 0 & -a & (1+a^2) & -a & \cdots & 0 \\ \vdots & & & & & \\ 0 & \cdots & 0 & -a & (1+a^2) & -a \\ 0 & \cdots & & 0 & -a & 1 \end{bmatrix}$$

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.