B Sparse conditional probability

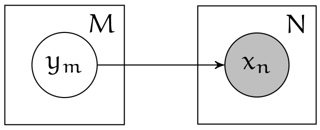

Figure 1: Generative diagram in plate notation.

Say we have a vector of events, $\mathbf{x} = (x_1, x_2, \ldots, x_N)^\mathsf{T}$, that is dependent upon another vector of events $\mathbf{y} = (y_1, y_2, \ldots, y_M)^\mathsf{T}$. We can write

$$p(x_n) = \sum_{m=1}^{M} p(x_n \mid y_m)\, p(y_m)\,. \tag{1}$$

This can be written in matrix form as

$$\begin{pmatrix} p(x_1) \\ p(x_2) \\ \vdots \\ p(x_N) \end{pmatrix} = \begin{pmatrix} p(x_1 \mid y_1) & p(x_1 \mid y_2) & \cdots & p(x_1 \mid y_M) \\ p(x_2 \mid y_1) & p(x_2 \mid y_2) & \cdots & p(x_2 \mid y_M) \\ \vdots & \vdots & \ddots & \vdots \\ p(x_N \mid y_1) & p(x_N \mid y_2) & \cdots & p(x_N \mid y_M) \end{pmatrix} \begin{pmatrix} p(y_1) \\ p(y_2) \\ \vdots \\ p(y_M) \end{pmatrix} \tag{2}$$

If $M \gg N$ then we have a sparse system.

If we now say that the events in $\mathbf{y}$ are mutually exclusive, then it can be a categorical distribution:

$$p(y_m) = p(y_m \mid \theta_m) = \theta_m, \qquad \sum_{m=0}^{M} \theta_m = 1. \tag{3}$$

If the events in $\mathbf{x}$ are also mutually exclusive, then all columns sum to unity. This may be important in training the matrix.

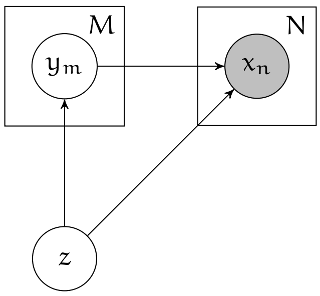

Figure 2: Generative diagram in plate notation for conditioning on z.

It's possible to condition the whole thing on something else, i.e.,

$$p(x_n \mid z) = \sum_{m=1}^{M} p(x_n \mid y_m, z)\, p(y_m \mid z)\,. \tag{4}$$

So,

$$\begin{pmatrix} p(x_1 \mid z) \\ p(x_2 \mid z) \\ \vdots \\ p(x_N \mid z) \end{pmatrix} = \begin{pmatrix} p(x_1 \mid y_1, z) & p(x_1 \mid y_2, z) & \cdots & p(x_1 \mid y_M, z) \\ p(x_2 \mid y_1, z) & p(x_2 \mid y_2, z) & \cdots & p(x_2 \mid y_M, z) \\ \vdots & \vdots & \ddots & \vdots \\ p(x_N \mid y_1, z) & p(x_N \mid y_2, z) & \cdots & p(x_N \mid y_M, z) \end{pmatrix} \begin{pmatrix} p(y_1 \mid z) \\ p(y_2 \mid z) \\ \vdots \\ p(y_M \mid z) \end{pmatrix} \tag{5}$$

C Manifolds

Given two numbers, x_1 and y_1, it is common to define a linear combination, w_1, using a variable λ, where $0 \leqslant \lambda \leqslant 1$:

$$w_1 = \lambda x_1 + (1 - \lambda) y_1. \tag{6}$$

This has the result that $x_1 \leqslant w_1 \leqslant y_1$. Adding in another variable, w_2, the situation can be written in vector form:

$$\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \lambda \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + (1 - \lambda) \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}. \tag{7}$$

This defines the red line on the left of figure 3.

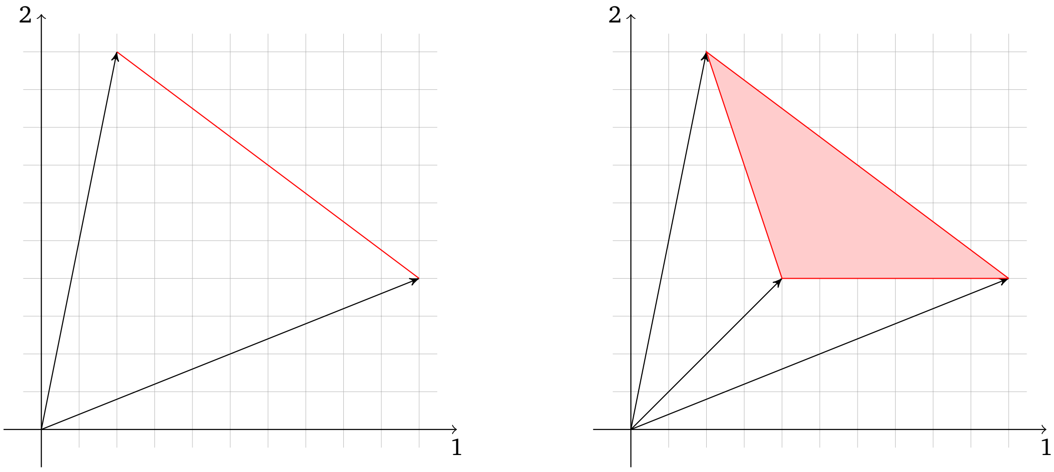

Figure 3: Left: Two vectors in 2D space. Right: Three vectors.

The situation can generalise to three vectors:

$$\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \lambda_1 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \lambda_2 \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \lambda_3 \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \qquad \sum_{i=1}^{3} \lambda_i = 1, \tag{8}$$

defining the shaded area on the right of figure 3. Notice that the shaded area is a simplex.

Generalising further to M vectors in $\mathbb{R}^N$, and rearranging,

$$\boldsymbol{w} = \mathbf{X}\boldsymbol{\lambda} \qquad \sum_{i=1}^{M} \lambda_i = 1, \tag{9}$$

where the M vectors are now the columns of $\mathbf{X}$.

Without the sum to one constraint, it's just a change of basis. The sum to one constrains $\boldsymbol{w}$ to lie within a simplex defined by the vectors.

Now imagine that the M vectors sample an S dimensional manifold in $\mathbb{R}^N$. If the space is well sampled, any point $\boldsymbol{w}$ *lying in the manifold* can be represented by $S + 1$ vectors. If $S \ll M$ then λ is sparse.