0.1 Bayesian Network

We model the state of each lymph node level (LNL) as a hidden or unobserved binary random variable, which indicates via values 0 or 1 if an LNL is healthy or involved, respectively. This state indicates if there is truly tumor present in an LNL, including the presence of occult metastases for the involved state – motivating the term hidden or unobserved state. Every LNL can be diagnosed using one or multiple modalities. Most used for diagnosis are imaging techniques like positron emission tomography (PET), computed tomography (CT) and magnetic resonance imaging (MRI), but palpation or fine needle aspiration (FNA) are also used. The diagnosis too, is modelled as binary random variable – this time an observed one – taking on 0 for negative and 1 for positive.

For notational convenience, we collect the hidden and observed random variables in a random vector each:

$$\begin{aligned} \text{hidden} \quad \mathbf{X} &= (X_v) \to \{0,1\}^V \\ \text{observed} \quad \mathbf{Z} &= (Z_v) \to \{0,1\}^V \end{aligned} \tag{1}$$

where V is the number of LNLs $v \in \{1, 2, \ldots, V\}$ in the graph. The conditional probabilities that link the hidden state to the observations can be written as follows:

$$\begin{aligned} P_{BN}\left(Z_v = z_v \mid X_v = x_v\right) = \left(z_v + (-1)^{z_v} \cdot s_P\right)(1 - x_v) \\ + \left((1 - z_v) + (-1)^{1-z_v} \cdot s_N\right) x_v \end{aligned} \tag{2}$$

with s_N and s_P being the sensitivity and specificity of the used diagnostic method. For example, for the probability of a false negative observation, i.e. diagnostic modality misses the presence of tumor, we get

$$P_{BN}\left(Z_v = 0 \mid X_v = 1\right) = 1 - s_N \tag{3}$$

Spread of the tumor through the lymphatic network is represented in this model by directed arcs to and between LNLs as illustrated in **??**. We introduce an additional vertex to the graph representing the primary tumor, which we assume to be the only one. Directed arcs from the primary tumor to an LNL represent direct spread of tumor cells from the primary tumor to the LNL. These arcs are associated with parameters b_v that we call base probabilities, and which indicate the probability that the tumor spreads directly to LNL v. When LNL s receives efferent lymphatics from LNL r, this too is represented by a directed arc from LNL r to s, and $r = \mathrm{pa}(s)$ which is called a parent node of s. These arcs are associated with a transition probability t_{rs} from r to s. The resulting directed acyclic graph (DAG) is shown in **??**, comprising ipsilateral levels I, II, II, through IV, and will be used throughout this work. However, when more data of detailed LNL involvement including additional levels becomes available and/or contralateral involvement, the model can be extended. The parameters b_v and t_{rs} associated with the directed arcs represent conditional probabilities, i.e. b_v answers the question given that all parent nodes are healthy, how likely is it that the primary tumor spreads to node v? t_{rs} on the other hand, can answer the question assuming no efferent spread from the primary tumor and given that all parent nodes except r are healthy, what is the likelihood of spread to node s? The conditional probability for involvement of LNL v given the state of its parent nodes is then given by

$$\begin{aligned} &P_{BN}\left(X_v = x_v \mid X_{\mathrm{pa}(v)} = x_{\mathrm{pa}(v)}, b_v, t_{\mathrm{pa}(v)v}\right) \\ &\qquad\qquad = x_v + (-1)^{x_v}(1-b_v)(1-t_{\mathrm{pa}(v)v})^{x_{\mathrm{pa}(v)}} \end{aligned} \tag{4}$$

We note here that this parametrization assumes the independence of causal influences (ICI), thereby allowing us to describe the model using only a few interpretable parameters. Dropping this assumption, a Bayesian network (BN) can also be defined using conditional probability tables (CPT) that have columns for every possible combinations of parent states. However, with the increase of the number of parent nodes (causes) in the graph, the number of parameters in the respective CPT would grow exponentially.

For the graph in **??** we can write down the parametrized CPT in the following manner:

$$\begin{aligned} &P_{BN}\left(X_v = 0 \mid X_{\mathrm{pa}(v)} = 0\right) = 1 - b_v \\ &P_{BN}\left(X_v = 1 \mid X_{\mathrm{pa}(v)} = 0\right) = b_v \\ &P_{BN}\left(X_v = 0 \mid X_{\mathrm{pa}(v)} = 1\right) = (1 - b_v)\left(1 - t_{\mathrm{pa}(v)v}\right) \\ &P_{BN}\left(X_v = 1 \mid X_{\mathrm{pa}(v)} = 1\right) = 1 - (1 - b_v)\left(1 - t_{\mathrm{pa}(v)v}\right) \end{aligned} \tag{5}$$

In case of a more general network, in which some LNLs receive efferent lymphatics from multiple other LNLs, eq. (5) can be generalised and the conditional probability of the hidden state becomes

$$\begin{aligned} &P_{BN}\Big(X_v = x_v \mid \left\{X_r = x_r\,,\, t_{rv}\right\}_{r\in\mathrm{pa}(v)}, b_v\Big) \\ &\qquad\qquad = x_v + (-1)^{x_v}(1-b_v)\prod_{r\in\mathrm{pa}(v)}(1-t_{rv})^{x_r} \end{aligned} \tag{6}$$

where we marginalized over all hidden variables X. Here we have assumed that each patient's diagnosis $\mathbf{z} = (z_1 \quad z_2 \quad \cdots \quad z_V)$ is complete, meaning that we have a diagnosis for each LNL. The likelihood can then be used to infer the model parameters via maximum likelihood inference or sampling.