6. Representing HMMs as exponential families

Definition 1

An exponential family of distributions for a random variable $X \in \mathcal{X}$ is a parametric model with densities

$$p_\theta(x) = h(x) \exp\left[\langle \varphi(x), \theta \rangle - a(\theta)\right]$$

where

- $\theta \in \mathbb{R}^n$ is the natural parameter
- $\varphi(x) \in \mathbb{R}^n$ is the sufficient statistics
- $h(x) \geq 0$ is the base measure
- $a(\theta)$ is the log partition function (cumulant function) given by

$$a(\theta) = \log \int h(x) \exp\langle \varphi(x), \theta \rangle \, d\nu(x). \qquad \square$$

Example 1

a) Bernoulli distribution $p_\theta(x) = \theta^x (1-\theta)^{1-x}, \quad x = 0, 1$

$$p_\theta(x) = \exp\left[x \log \frac{\theta}{1-\theta} + \log(1-\theta)\right]$$

b) Univariate normal distribution $p_\mu(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-\mu)^2\right]$

$$h(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

$$\varphi(x) = x$$

$$a(\mu) = \frac{1}{2}\mu^2$$

Minimal representation:

- $\nexists\ b \in \mathbb{R}^n : \quad \langle b, \varphi(x) \rangle = \text{const} \quad \forall x \in \mathcal{X}$
- $\nexists\ b \in \mathbb{R}^n : \quad \langle b, \theta \rangle = \text{const} \quad \forall \theta \in \Theta$

Def 16, Sec. 1 $\Rightarrow$ the joint p.d. of a Markov chain model with strictly positive prob's can be written as

$$p(s) = p(s_1, .., s_n) = \frac{1}{Z} \prod_{i=2}^{n} g_i(s_{i-1}, s_i) = \frac{1}{Z} \exp \sum_{i=2}^{n} u_i(s_{i-1}, s_i)$$

Remark 1 The factors g_i, resp. the potentials u_i define the model uniquely. The reverse is not true.

Remark 2 The partition function $Z(u)$ is defined by

$$Z(u) = \sum_{s \in K^n} \exp \sum_{i=2}^{n} u_i(s_{i-1}, s_i)$$

and can be computed by an algorithm similar to the one discussed in Sec. 3.

Denote:

(1) $\varphi(s_i) \in \mathbb{R}^K$ the binary valued indicator vector that denotes the state $s_i \in K$ in „one out of K" encoding, i.e.

$$\varphi(s_i = k) = (0, .., 1, ... 0)$$

(2) U_i the $K \times K$ matrix with values $u_i(s_{i-1}, s_i)$

The joint p.d. of a strictly positive Markov chain model can be written as

$$p(s) = \frac{1}{Z(u)} \exp \sum_{i=2}^{n} \langle \varphi(s_{i-1}), u_i \varphi(s_i) \rangle$$
$$= \frac{1}{Z(u)} \exp \sum_{i=2}^{n} \langle \Phi(s_{i-1}, s_i), u_i \rangle,$$

where

$$\Phi(s_{i-1}, s_i) = \varphi(s_{i-1}) \otimes \varphi(s_i)$$

is a $K \times K$ binary valued indicator matrix and

$$\langle \Phi, u \rangle = \mathrm{Tr}(\Phi^T u)$$

denotes the Frobenius inner product.

Finally, denote $\Phi = (\Phi_2, \ldots, \Phi_n)$ and $u = (u_2, \ldots, u_n)$ and write

$$p(s) = \frac{1}{Z(u)} \exp \langle \Phi(s), u \rangle$$

The joint p.d. of an HMM can be written as

$$p(s) = \frac{1}{Z(u)} \exp \langle \Phi(x, s), u \rangle$$

by using similar notations.

<u>Remark 3</u> The EF-representations of Markov models / HMMS are not minimal.

<u>Remark 4</u> The components of the expectation

$$\mathbb{E}_{s \sim p_u(s)} [\Phi(s)] \quad \text{for a Markov chain model}$$

are the pairwise marginal probabilities for pairs of ~~conseq~~ consecutive states.

7. ML estimator for supervised learning of HMMS

Given an i.i.d. sample of pairs of sequences

$$T = \{(x^j, s^j) \mid x^j \in F^n, \; s^j \in K^n, \; j = 1, \dots, \ell\}$$

estimate the model parameters of the HMM by the maximum likelihood estimator

$$u^* \in \arg\max_u \prod_{(x,s) \in T} p_u(x,s) =$$

$$= \arg\max_u \frac{1}{|T|} \sum_{(x,s) \in T} \log p_u(x,s),$$

i.e. find optimal $\tilde{u}_i^*(x_i, s_i)$, $u_i^*(s_{i-1}, s_i)$ or, equivalently, $p(x_i, s_i)$, $p(s_{i-1}, s_i)$.

<u>Intuitive answer</u>: u^* is given by

$$p_{u^*}(s_{i-1}, s_i) = \beta(s_{i-1}, s_i)$$

$$p_{u^*}(x_i, s_i) = \beta(x_i, s_i)$$

where β-s denote the frequencies of the corresponding events in T.

Let us prove correctness. The log-likelihood of T is

$$L(u) = \frac{1}{|T|} \sum_{(x,s) \in T} \left[\langle \varphi(x,s), u \rangle - \log Z(u) \right]$$

$$= \langle \psi, u \rangle - \log Z(u),$$

where

$$\psi = \mathbb{E}_T \varphi = \frac{1}{|T|} \sum_{(x,s) \in T} \varphi(x,s)$$

<u>Remark 1</u> Observe that all we need to know from the sample T is $\psi = \mathbb{E}_T \varphi$

<u>Lemma 1</u> The log-partition function $\log Z(u)$ of an HMM is convex in u.

<u>Proof</u>

$$\nabla_u \log Z(u) = \frac{1}{Z(u)} \sum_{x,s}{}' \exp\langle \varphi(x,s), u\rangle \, \varphi(x,s) \overset{!}{=} \mathbb{E}_u \varphi$$

Recall that the components of $\mathbb{E}_u \varphi$ are the pairwise marginal prob's on the edges of the model.

$$\nabla_u^2 \log Z(u) = \mathbb{E}_u [\varphi \otimes \varphi] - \mathbb{E}_u[\varphi] \otimes \mathbb{E}_u[\varphi]$$

$$= \mathbb{E}_u \big[(\varphi - \mathbb{E}_u \varphi) \otimes (\varphi - \mathbb{E}_u \varphi) \big]$$

The expectation of a positive semidefinite matrix is p.s.d. $\Rightarrow$ $\log Z(u)$ is convex. $\square$

The log-likelihood is concave and has global maxima only as a consequence. They are given by

$$\nabla_u L(u^*) = \frac{1}{|T|} \sum_{(x,s)\in T}{}' \varphi(x,s) - \mathbb{E}_{u^*}[\varphi] = \mathbb{E}_T[\varphi] - \mathbb{E}_{u^*}[\varphi] = 0$$

Recall that the components of $\mathbb{E}_u[\varphi]$ are the pairwise marginal prob's of the model $p_u(x,s)$. Hence, the optimiser u^* defines the model whose pairwise marginal prob's coincide with the empirical marginal frequencies in T.