Supervised learning of GRFs

A. Maximum likelihood estimators

Given: A parametrised family of p.d.s $p_u(x)$, $x \in \mathcal{X}$, $u \in \mathcal{U}$ and a sample $T^m = \{x^j \mid j = 1, \dots, m\}$, where x^j are generated i.i.d. from $p_{u_0}(x)$ with unknown $u_0 \in \mathcal{U}$

Task: Estimate the unknown parameter u_0

Maximum likelihood estimator

$$u_* \in \operatorname*{argmax}_{u \in \mathcal{U}} \frac{1}{m} \sum_{j=1}^{m} \log p_u(x^j) = \operatorname*{argmax}_{u \in \mathcal{U}} L(u, T^m)$$

Consistency of MLEs

$$L(u, T^m) \xrightarrow[m \to \infty]{P} L(u) = \sum_{x \in \mathcal{X}} p_{u_0}(x) \log p_u(x)$$

$$\operatorname*{argmax}_{u \in \mathcal{U}} \downarrow \qquad\qquad \downarrow \operatorname*{argmax}_{u \in \mathcal{U}}$$

$$u_*(T^m) \xrightarrow[m \to \infty]{P} u_0$$

a) $L(u, T^m) \xrightarrow[m \to \infty]{P} L(u)$? Yes, LLN

$$\mathbb{P}\left(|L(u, T^m) - L(u)| \geq \varepsilon \right) \xrightarrow[m \to \infty]{} 0 \quad \forall u \in \mathcal{U}, \forall \varepsilon > 0$$

b) $u_0 = \operatorname*{argmax}_{u \in \mathcal{U}} L(u)$? Yes, if the model is identifiable, i.e. $p_{u_0}(x) \not\equiv p_u(x)$ $\forall u \neq u_0$. This follows from $D_{KL}(p \| q) \geq 0$ with equality iff $p = q$

c) $u_*(T^m) \xrightarrow[m\to\infty]{P} u_0$? Yes, if

1) $\mathcal{U}$ is compact and $L(u, T^m)$ is continuous in u.

2) $L(u, T^m)$ converges <u>uniformly</u>, i.e.

$$P\left(\sup_{u\in\mathcal{U}} |L(u, T^m) - L(u)| \geq \varepsilon\right) \xrightarrow[m\to\infty]{} 0$$

<u>Remark 1</u> Conditions 1) & 2) are sufficient but not necessary. E.g. condition 1) can be replaced by $L(u, T^m)$ is concave in u and $u_0 \in \text{int}(\mathcal{U})$.

B. <u>MLEs for Gibbs random fields</u>

$S = \{S_i \mid i \in V\}$ is a GRF w.r.t. the graph (V, E) and has p.d.

$$p_u(s) = \frac{1}{Z(u)} \exp \sum_{ij\in E} u_{ij}(s_i, s_j).$$

The parameters u_{ij} are unknown. We are given an i.i.d sample of training data $T^m = \{s^j \in K^V \mid j = 1, \dots, m\}$.

Since the GRF is an exponential family, we can write

$$L(u, T^m) = \frac{1}{m} \sum_{s^j \in T^m} \log p_u(s^j) =$$

$$= \frac{1}{m} \sum_{j=1}^{m} \langle \Phi(s^j), u\rangle - \log Z(u)$$

- $L(u, T^m)$ is concave in u
- The model is identifiable up to re-parametrisations
- If $\mathcal{U}$ is compact, then the MLE is consistent

How to solve the task

$$L(u, T^m) = \frac{1}{m} \sum_{j=1}^{m} \langle \varphi(s^j), u \rangle - \log Z(u) \to \max_{u \in \mathcal{U}}$$

for a GRF? Gradient ascend requires to compute

$$\nabla L(u, T^m) = \frac{1}{m} \sum_{j=1}^{m} \varphi(s^j) - \mathbb{E}_u[\varphi] .$$

Computing $\mathbb{E}_u[\varphi]$ amounts to computing pairwise marginals for all edges of the graph. Notice, that the marginals must be (re-) estimated in each iteration of the gradient ascend.

L6 Pseudo-likelihood estimators for GRFs

Can we do simpler? Besag, 1975 $\Rightarrow$ Recall that a GRF on a graph (V, E) is defined by fixing the family of conditional distr.

$$p_u(s_i \mid s_{N_i}) = \frac{1}{Z_i(u, s_{N_i})} \exp \sum_{j \in N_i} u_{ij}(s_i, s_j)$$

(see Gibbs sampler). Hence, we may use pseudo-likelihood

$$\tilde{L}(u, T^m) = \frac{1}{m} \sum_{s \in T^m} \sum_{i \in V} \log p_u(s_i \mid s_{N_i}) \to \max_u$$

instead of MLE. We obtain

$$\tilde{L}(u, T^m) = \frac{1}{m} \sum_{s \in T^m} \sum_{i \in V} \sum_{j \in N_i} u_{ij}(s_i, s_j) - \frac{1}{m} \sum_{s \in T^m} \sum_{i \in V} \log Z_i(u, s_{N_i})$$

$$= 2 \sum_{ij \in E} \frac{1}{m} \sum_{s \in T^m} u_{ij}(s_i, s_j) - \sum_{i \in V} \frac{1}{m} \sum_{s \in T^m} \log Z_i(u, s_{N_i})$$

- Computing $\tilde{L}(u, T^m)$ and $\nabla \tilde{L}(u, T^m)$ has complexity $O(m |E| |K|^2)$

- $\tilde{L}(\eta, T^m)$ is concave. It can be proved that pseudo-likelihood estimators are consistent.
- Its variance is higher than the variance of MLE
- Pseudo-likelihood estimators can not be used for unsupervised learning, whereas MLE can (by EM-algorithm).