Supervised learning of GRFs

1A. ML-estimator for Gibbs random fields

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

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

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

MLE

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

Since the GRF is an exponential family, we can write

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

- $L(u, \mathcal{T}^m)$ is concave in u
- The model is identifiable up to re-parametrisations
  I.e. $L(u) = \mathbb{E}_{\mathcal{T}^m \sim p_{u_0}} \big(\log p_u(\mathcal{T}^m) \big)$ has maximum at u_0 up to re-parametrisations
- If $\mathcal{U}$ is compact, then the MLE is consistent.

How to solve the task

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

Computing the gradient

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

amounts to computing pairwise marginals for all edges of the graph!

B. Pseudo-likelihood estimator for GRFs

Can we do simpler? Besag, 1995 ⇒ Recall that a GRF on a graph (V, E) is defined by fixing the (consistent) family of conditional distributions

$$p_u(s_i \mid s_{\mathcal{N}_i}) = \frac{1}{Z_i(u, s_{\mathcal{N}_i})} \exp \sum_{j \in \mathcal{N}_i} u_{ij}(s_i, s_j) \qquad \forall i \in V$$

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

$$\tilde{L}(u, T^m) = \frac{1}{m} \sum_{s \in T^m} \sum_i \log p_u(s_i \mid s_{\mathcal{N}_i}) \to \max_{u \in \mathcal{U}}$$

instead of MLE. We obtain

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

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

Computing $\tilde{L}(u, T^m)$ and $\nabla \tilde{L}(u, T^m)$ has complexity $\mathcal{O}(m |E| |K|^2)$. Properties of pseudo-likelihood estimators:

(1) $\tilde{L}(u, T^m)$ is concave in u.

(2) pseudo-likelihood estimators are consistent

(3) their variance (in T^m) is higher than the variance of MLE

(4) PLE can not be used for unsupervised learning
(MLE + EM algorithm allows unsupervised learning)

Example 1 Conditional random fields for semantic segmentation

Combine a convolutional neural network with a CRF

Denote: $D \subset \mathbb{Z}^2$ – image domain

$x: D \to \mathbb{R}^3$ – image

$s: D \to K$ – segmentation

CNN network $a = f(x, w)$, $a \in \mathbb{R}^{DK}$ computes features for each pixel $i \in D$ and each segment label $k \in K$.

This could be e.g. a U-net

CRF model:

$$p_\theta(s \mid x) = \frac{1}{Z(x, \theta)} \exp\Big[\sum_{ij \in E} u_{ij}(s_i, s_j) + \sum_{i \in V} f_{s_i}(x, w)\Big]$$

where $\theta = (u, w)$. Such models can be learned end-to-end by PLE. □