Pseudo-Codes and Convergence Guarantee: Automatic Discovery of Disease Subtypes by Contrasting with Healthy Controls

Robin Louiset, Edouard Duchesnay, Benoit Dufumier, Antoine Grigis, Pietro Gori

1 SINKHORN-KNOPP SOFT K-MEANS

In this section, we describe the pseudo-code algorithm (See Alg. 1.) for the Soft K-Means algorithm regularized with Sinkhorn-Knopp [2]. We implement this algorithm on GPU. The Sinkhorn-Knopp algorithm directly comes from [1] and uses the same hyper-parameter choice.

Algorithm 1 SK regularized Soft K-Means pseudo-code

Input:
  Disease representations: $Z \in \mathbf{R}^{N_{y=1} \times D}$,
  K: subgroups number, λ: SK temperature
  N: iterations
Output:
  Centroids: $\mu = \{\mu^k\}_{k \in |[1,K]|}$.
Initialization step:
  Initialize centroids μ with K-Means ++ algorithm.
for i in N iterations **do**
  Compute soft clustering probabilities $Q(c_i)$ given a representation Z_i: $Q(c_i) = \frac{1/||Z_i - \mu_i||_2^2}{\sum_{j=1}^{K}(1/||Z_i - \mu_j||_2^2)}$.
  Apply SK regularization: $Q = SK(Q, \lambda)$.
  Compute one-hot clustering matrix Q_{hot}:
    $Q_{hot} = OneHot(Q.argmax(dim = 1))$.
  for k in K subgroups **do**
    Update centroid k: $\mu^k = \frac{(Z * Q_{hot}[:,k]).sum()}{Q_{hot}[:,k].sum()}$.
  end for
end for
Return centroids $\mu = \{\mu^k\}_{k \in |[1,K]|}$.

The OneHot(.) function consists of transforming a smooth probability clustering vector (e.g.: $[0.2, 0.1, 0.7]$) into the hard version (i.e.: $[0, 0, 1]$).

2 CLUSTERING RE-IDENTIFICATION

In the clustering re-identification paragraph, our objective is to identify each updated cluster (epoch $t+1$) with its most similar previous cluster (epoch t). Let us clarify the notation. At epoch t, we have estimated K subtypes, we can compute their respective centroids with the following formula:

$$\mu_k^t = \sum_{i=1}^{N} 1_{c_i^t = k} f_\theta(x_i) \tag{1}$$

where f_θ is the encoder, x_i is an input image, associated with an inferred C^t at epoch t.
At epoch $t+1$, we update our subtype estimation, and we estimate K updated subtypes, once again, we can compute their centroids:

$$\mu_k^{t+1} = \sum_{i=1}^{N} 1_{c_i^{t+1} = k} f_\theta(x_i) \tag{2}$$

We wish to permute the labels of the clusters (and their centroids) estimated at epoch $t+1$ so that there is a continuity between clusters estimated at epoch t and those estimated at epoch $t+1$. In practice, we aim to compute a permutation function σ that maps an updated cluster (epoch $t+1$) onto its most similar former cluster (epoch t). Given a similarity function $s(\mu, \mu')$ between two centroids μ and μ'. We are seeking the optimal permutation σ^*, which maximizes the average similarity:

$$\sigma^* = \max_{\sigma} \sum_{k=1}^{K} s(\mu_k^t, \mu_{\sigma^{-1}(k)}^{t+1}) \tag{3}$$

Importantly, we wish to construct a function σ that is bijective. Indeed, as explained in the main text, a non-bijective mapping could potentially allow for more than one previous cluster to be merged into a single updated cluster, which may produce one or more empty clusters. For example, assuming that $K = 2$ and that the estimated mapping gives $\sigma(0) = 1$, $\sigma(1) = 1$, then after having permuted the indices of the updated clusters, we would get $C_0^{t+1} = \varnothing$ because $\sigma^{-1}(0) = \varnothing$). Thus, to ensure the bijectivity of σ, we propose casting our problem into a conceptually different one. Let us explain it in detail.
Let assume that we are given K data-points: $\{c_j^{t+1}, j \in$

$|[1, K]|\}$ (in our experiment, it corresponds to the K centroids of clusters estimated at epoch $t+1$). Now, let's say that we are given K categories (which, in our case, correspond to the K clusters estimated at epoch t). Given a similarity measure, the probability of a sample j to belong to a given category i can be computed with the following formula:

$$p(c_i^t|\mu_j^{t+1}) = \frac{s(\mu_j^{t+1}, \mu_i^t)}{\sum_{k=1}^{K} s(\mu_j^{t+1}, \mu_k^t)} \tag{4}$$

We wish to find the closest solution where the samples get assigned to a category, and each category has the same number of attributed samples (equipartition property). This problem has a simple solution that can be easily estimated via an optimal transport algorithm: the Sinkhorn-Knopp algorithm. See Alg. 2.

Importantly, note that in our case, as we have K samples for K classes, the equipartition property is respected if and only if each sample gets mapped to a single category, which is equivalent to having a bijective mapping between samples and categories.

Algorithm 2 Subgroups re-identification pseudo-code

```
Inputs:
    K: subgroups number
    Previous Subgroups Centroids: μ^t = {μ_k^t}_{k∈|[1,K]|}
    Subgroups Centroids: μ^{t+1} = {μ_k^{t+1}}_{k∈|[1,K]|}
Output:
    Permuted Subgroups Centroids: μ^{t+1} = {μ_{σ^{-1}(k)}^{t+1}}_k
Initialization step: Compute the similarity matrix S:
    S = (μ^t / ||μ^t||_2)^T . (μ^{t+1} / ||μ^{t+1}||_2)
while len(np.unique(σ)) ≤ K
    Apply SK regularization: S_SK = SK(S, λ)
    Compute permutation: σ = np.argmax(S_SK, axis=1)
    Increase SK strength: λ = 1.1 × λ
endwhile
Return permuted centroids μ^{t+1} = μ^{t+1}[σ, :]
```

3 Convergence guarantee

Here, we provide proof that the proposed Expectation-Maximization optimization process yields a monotonic increase of the log of the joint conditional likelihood. The proof is very similar to the one proposed in [**?**]. Calling $F(\theta, \phi, \psi)$ the joint conditional likelihood, namely our cost function (Eq. 2 in the main text), we have:

$$\begin{aligned} F(\theta, \phi, \psi) &= \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} Q(c_i = k) \frac{p_{\theta,\phi,\psi}(y_i, c_i = k|x_i)}{Q(c_i = k)} \right) \\ &\geq \sum_{i=1}^{n} \sum_{k=1}^{K} Q(c_i = k) \log p_{\theta,\phi}(y_i|c_i = k, x_i) \\ &\quad - D_{KL}(Q(c)||p_{\theta,\psi}(c|x)) \end{aligned} \tag{8}$$

Given a guess of the parameters $\theta^{(t)}$ at the t-th step, the E-step consists in choosing $Q^{(t)} = p_{\theta^{(t)}}(c_i|x_i, y_i)$ which makes the previous bound (Eq. 8) tight (*i.e.,* the inequality holds with equality). This means that, with this choice of $Q^{(t)}$, we have:

$$\begin{aligned} &F(\theta^{(t)}, \phi^{(t)}, \psi^{(t)}) = \\ &\sum_{i=1}^{n} \sum_{k=1}^{K} Q^{(t)}(c_i = k) \log p_{\theta^{(t)},\phi^{(t)}}(y_i|c_i = k, x_i) \\ &- D_{KL}(Q(c)||p_{\theta^{(t)},\psi^{(t)}}(c|x)) \end{aligned} \tag{9}$$

At the t-th M-step, we freeze $Q^{(t)}$ and we obtain the parameters $\theta^{(t+1)}$, $\psi^{(t+1)}$ and $\phi^{(t+1)}$ by maximizing the right-hand side of the equation above (Eq. 5 in the main text). Thus:

$$\begin{aligned} F(\theta^{(t+1)}, &\phi^{(t+1)}, \psi^{(t+1)}) \geq \\ &\sum_{i=1}^{n} \sum_{k=1}^{K} Q^{(t)}(c_i = k) \log p_{\theta^{(t+1)},\phi^{(t+1)}}(y_i|c_i = k, x_i) \\ &- D_{KL}(Q^{(t)}||p_{\theta^{(t+1)},\psi^{(t+1)}}(c|x)) \geq \\ &\sum_{i=1}^{n} \sum_{k=1}^{K} Q^{(t)}(c_i = k) \log p_{\theta^{(t)},\phi^{(t)}}(y_i|c_i, x_i) \\ &- D_{KL}(Q^{(t)}||p_{\theta^{(t)},\psi^{(t)}}(c|x)) = F(\theta^{(t)}, \phi^{(t)}, \psi^{(t)}) \end{aligned} \tag{10}$$

where the first inequality comes from Eq. 8 and the second one is true since we look for the parameters $\theta^{(t+1)}, \phi^{(t+1)}, \psi^{(t+1)}$ that maximizes $F(\theta^{(t)}, \phi^{(t)}, \psi^{(t)})$. The above result suggests that $F(\theta, \phi, \psi)$ monotonically increases.

References

[1] Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. NeurIPS **33**, 9912–9924 (2020)

[2] Cuturi, M.: Sinkhorn Distances: Lightspeed Computation of Optimal Transport p. 9